Croisements,ordres et ultrametriques:Application a la recherche de concensus en classification automatique

Edwin Diday

INRIA

ENTRE DE ROCQUENCOURT

Institut National
de Recherche
en Informatique
et en Automatique

Domaine de Voluceau
Rocquencourt
BP105
78153 Le Chesnay Cedex
France
Tél. 954 90 20

Rapports de Recherche

N° 144

CROISEMENTS, ORDRES ET ULTRAMÉTRIQUES : APPLICATION À LA RECHERCHE DE CONSENSUS EN CLASSIFICATION AUTOMATIQUE

Edwin DIDAY

Juillet 1982

CROISEMENTS, ORDRES ET ULTRAMETRIQUES : APPLICATION A LA RECHERCHE DE CONSENSUS EN CLASSIFICATION AUTOMATIQUE

Edwin DIDAY

Résumé

La représentation visuelle d'une hiérarchie induit un ordre sur les singletons. Si l'on désire représenter la même hiérarchie en tenant compte de contraintes extérieures (ordre des singletons induit par une autre hiérarchie, une partition, un indice de dissimilarité, par exemple) des croisements peuvent apparaître. Il y a un croisement dans la représentation visuelle d'une hiérarchie quand une branche horizontale (associée à un palier) est coupée par une branche verticale associée à un singleton.

On utilise la notion de compatibilité entre un ordre et un indice de dissimilarité ; on introduit les notions de semi-compatibilité et compatibilité faible. On étudie les aspects matriciels qui débouchent sur une généralisation des matrices de Robinson. On fait le lien entre toutes ces notions et les chaînes de longueur minimales au sens de l'indice de dissimilarité choisi. Dans le cas où cet indice est une ultramétrique, on obtient des propriétés intéressantes liant l'ordre des singletons correspondant à la visualisation d'une hiérarchie indicée et l'ultramétrique induite par cette hiérarchie. Tous ces résultats débouchent sur des algorithmes nouveaux permettant de construire des hiérarchies adaptatives (sous contraintes d'ordre, de distance, de partition),des hiérarchies qui soient visuellement proches (en réajustant l'ordre des singletons) ou une hiérarchie qui réalise un "consensus" entre plusieurs hiérarchies, etc...

Abstract

One of the most important and difficult problems encountered in automatic classification is that of comparison of classifications. The problem arises when we want to compare the same set of objects characterized by several data arrays. For instance, a time-series of data arrays or data arrays each depending on a different set of variables. This problem also arises when we wish to study the effect of different coding transformations, different choices of dissimilarity indices, the robustness of classification obtained, etc. The notion of crossing sheds new light in this framework. It allows us to relate the visual representation of a hierarchy and the notion of compatibility between an order and a dissimilarity index for which matrix characterizations are provided. The theoretical results provide simple and effective algorithms which facilitate the visual comparison of classifications and the study of consensus between them.

Keywords

Clustering, hierarchies, ultrametrics, orders, consensus.

TABLES DES MATIERES

1. - INTRODUCTION

Parmi les problèmes qui se posent dans la pratique de la classification automatique, l'un des plus fréquents et difficiles est celui de la comparaison de différentes structures inter-classes (hiérarchies, partitions, recouvrements, etc.). Ce problème se pose quand on désire comparer une même population d'individus caractérisés par des tableaux de données différents : tableaux évoluant dans le temps, tableaux correspondants à différents paquets de variables, par exemple. Il se pose aussi quand on désire comparer l'effet d'un changement de codage, le choix de différentes mesures de ressemblance, la robustesse de la structure inter-classe choisie pour les données traitées, etc.

Les retombées pratiques de tous ces problèmes ont attiré l'attention de nombreux chercheurs ; citons Adams (1972) qui s'intéresse principalement à la comparaison d'arbres et de hiérarchies par recherche de "consensus" (chercher, par exemple, la hiérarchie qui "ressemble" le plus à plusieurs hiérarchies), Farris (1973) et Minkowitz (1978) qui s'intéressent à la définition de mesures de ressemblance entre hiérarchies, Hubert (1977) pour les aspects "robustesse" ; citons surtout Rohlf (1981) qui fait la synthèse des approches les plus récentes.

La notion de "croisement" donne un éclairage nouveau à toute cette problématique. Elle nous est d'abord apparueen classification hiérarchique (les propriétés qui la caractérise ont été ensuite étendues au cas d'indices de dissimilarité autres que des ultramétriques) ; la classification hiérarchique fournit une structure inter-classe qui est très utilisée car d'interprétation visuelle facile quand la taille de la population n'est pas trop grande ; par contre, la comparaison visuelle de plusieurs hiérarchies n'est pas très aisée surtout si l'ordre des singletons qui se trouvent à la base de chaque hiérarchie n'est pas le même. Si l'on tente de représenter des hiérarchies avec le même ordre sur les singletons, des "croisements" peuvent apparaître ; on dit qu'il y a un "croisement" dans la représentation visuelle d'une hiérarchie quand il apparaît une coupure entre une branche horizontale et une branche verticale associée à un singleton.

L'étude théorique de la notion de croisement débouche rapidement sur la recherche de liens entre un indice de dissimilarité et un ordre. On introduit la notion de "compatibilité faible" entre un ordre Θ et un indice de dissimilarité d et on montre que cette notion est équivalente, si d est une ultramétrique, à l'inexistence de croisements pour la hiérarchie induite par d quand Θ est l'ordre des singletons ; on montre aussi qu'elle est équivalente au fait que la chaîne induite par Θ et valuée par d soit de plus courte longueur ; quand d n'est pas une ultramétrique, il faut introduire une nouvelle condition appelée "semi-compatibilité" qui est plus restrictive que celle de "compatibilité faible" mais moins restrictive que celle de "compatibilité" qui a été clairement définie par Brossier (1981).

Les aspects matriciels de ces notions permettent de les relier entre elles et de généraliser les matrices de Robinson utilisée, notamment, par Kendall (1969) et Hubert (1974) à des familles de matrices dites SDR et SDD qui les contiennent. La semi-compatibilité donne un point de vue nouveau à la notion de chaîne T-minimax introduite par Leclerc (1974) et permet donc de faire apparaître une nouvelle caractérisation des chaînes "incluses" dans un arbre de longueur minimum. Il résulte de ces résultats théoriques des algorithmes permettant de construire des hiérarchies qui soient visuellement proches (en réajustant l'ordre des singletons) ou de construire une hiérarchie qui réalise un "consensus" entre plusieurs hiérarchies ; on propose également des algorithmes permettant la construction rapide et séquentielle d'une hiérarchie quand on connaît l'ultramétrique associée, la construction adaptative de hiérarchies sous contrainte de distance ou de partition, la recherche d'une chaîne qui soit dans un arbre de longueur minimum, la déformation minimum à donner à un tableau de distances pour qu'une chaîne soit un arbre de longueur minimum, la recherche d'anomalies entre hiérarchies ; pour comparer des hiérarchies on peut utiliser différents types de mesures de ressemblance, voir Rohlf (1981) par exemple ; on propose des mesures de ressemblance entre hiérarchies basées sur la notion de croisement.

Nous supposons connus un certain nombre de notions classiques en classification automatique (hiérarchie indicée, indice d'agrégation, ultramétriques, existence d'un bijection entre les hiérarchies indicées et les ultramétriques, etc.) que le lecteur pourra trouver dans le chapitre 2 du livre "Eléments d'analyse des donnée" Dunod (1982) ; signalons enfin que la notion de croisement correspond à une "inversion horizontale" dans une hiérarchie, le lecteur intéressé par les problèmes "d'inversion verticale" pourra se reporter à [5].

2. - DEFINITION D'UN CROISEMENT

Soit une hiérarchie H définie sur un ensemble d'individus Ω et un ordre Θ quelconque sur Ω que nous notons pour simplifier $w_1, \ldots, w_n$ (autrement dit, $w_i \ \Theta \ w_j$ si $i < j$).

Définition

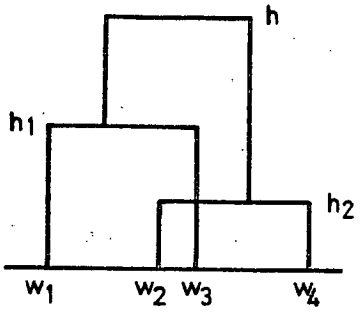
On dit que l'ordre Θ donne lieu à un croisement pour la hiérarchie H, si et seulement si il existe $h \in H$ contenant deux éléments de Ω w_{i_1} et w_{i_3} tels qu'il existe :

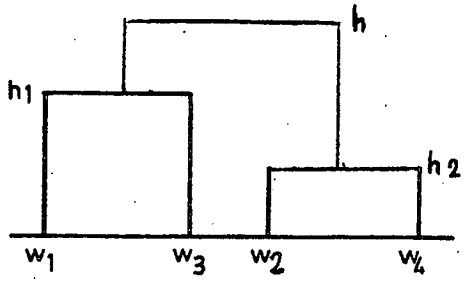
$$w_{i_2} \notin h \text{ avec } i_1 < i_2 < i_3$$

Exemples

L'ensemble des individus est $\Omega = \{w_1, w_2, w_3, w_4\}$, $h_1 = \{w_1, w_3\}$ et $h_2 = \{w_2, w_4\}$ dans la hiérarchie H qui est représentée figures 1 et 2.

Dans la figure 1, l'ordre $w_1\ w_2\ w_3\ w_4$ donne lieu à un croisement ; par contre, dans la figure 2, l'ordre $w_1\ w_3\ w_2\ w_4$ ne donne pas lieu à un croisement pour la même hiérarchie H.





3. - ORDRES SANS CROISEMENTS

3.1. - Construction d'un ordre sans croisement

On peut définir un ordre sans croisement à partir d'une hiérarchie H en associant, à chaque palier, un ordre sur les plus grands paliers dont il est la réunion ; en procédant ainsi à partir du palier identique à Ω jusqu'aux paliers qui ne contiennent que des singletons, on obtient un

ordre sur les individus qui ne peut comporter de croisements, par construction même. Remarquons que si H est une hiérarchie binaire (il y a donc n-1 paliers) alors le nombre d'ordres sans croisement est 2^{n-1} puisque chaque palier peut ordonner de deux façons les deux paliers dont il est la réunion. C'est beaucoup mais c'est peu par rapport au nombre de hiérarchies binaires sur Ω qui vaut $n!\ (n-1)!\ 2^{1-n}$ d'après Frank et Svensson (1981).

Exemple

Si l'on reprend la hiérarchie représentée dans la figure 2, l'algorithme indiqué ci-dessus donne la succession d'ordres suivants :

$$h \to h_1\ h_2,\ h_1 \to w_1\ w_3,\ h_2 \to w_2\ w_4$$

d'où finalement l'ordre $w_1\ w_3\ w_2\ w_4$.

3.2. - Une condition nécessaire et suffisante pour avoir un ordre sans croisement

Nous considérons l'ordre Θ sur Ω défini par $w_1\ w_2\ \ldots\ w_n$ et un indice de dissimilarité d.

Définition

Nous dirons que d et Θ sont faiblement compatibles si et seulement si, pour tout triplet, $w_{i_1}\ w_{i_2}\ w_{i_3}$ où $i_1 < i_2 < i_3$ et deux individus sont consécutifs (au sens de Θ), la distance (au sens de d) des deux sommets consécutifs est inférieure à $d(w_{i_1}, w_{i_3})$.

Autrement dit si $i_2 = i_1 + 1$ par exemple, on doit avoir $d(w_{i_1}, w_{i_2}) \leq d(w_{i_1}, w_{i_3})$.

On sait qu'il existe une bijection entre l'ensemble des hiérarchies indicées et l'ensemble des ultramétriques (voir, par exemple, E. DIDAY et al (1982)) ; soit δ_H l'ultramétrique associée à une hiérarchie indicée notée H par cette bijection ($\delta_H(w_i, w_j)$ est la hauteur du plus bas des paliers de H contenant w_i et w_j). On a alors la proposition suivante :

Proposition 1

Une c.n.s. pour que l'ordre Θ ne donne pas lieu à un croisement pour la hiérarchie indicée H est que δ_H et Θ soient faiblement compatibles

Démonstration

Montrons d'abord que s'il n'y a pas de croisement alors δ_H et Θ sont nécessairement faiblement compatibles ; en effet, si δ_H et Θ ne sont pas faiblement compatibles, alors il existe un triplet w_{i_1} w_{i_2} w_{i_3} tel que $\delta_H(w_{i_1}, w_{i_2}) > \delta_H(w_{i_1}, w_{i_3})$ si $i_2 = i_1 + 1$ ou $\delta_H(w_{i_2},w_{i_3}) > \delta_H(w_{i_1},w_{i_3})$ si $i_3 = i_2 + 1$; supposons que $i_2 = i_1 + 1$ (la démonstration se fait de façon analogue si $i_3 = i_2 + 1$). Soit h_2 (resp. h_3) le plus bas palier de H qui contienne w_{i_1} et w_{i_2} (resp. w_{i_3}) ; si w_{i_2} appartenait à h_3, on aurait $h_2 \subset h_3$ ou $h_2 \equiv h_3$ et donc $\delta_H(w_{i_1},w_{i_2}) \leq \delta_H(w_{i_1},w_{i_3})$ puisque H est indicée (i.e. pas d'inversions), ce qui est contraire à l'hypothèse ; donc w_{i_2} n'appartient pas à h_3 bien que $i_1 < i_2 < i_3$, il y a donc un croisement.

Réciproquement, si l'ordre Θ est sujet à un croisement pour la hiérarchie H, il existe h_1 et h_2 dans H tels que h_1 contienne w_{i_1}, w_{i_3} et h_2 contienne $w_{i_2} \notin h_1$ avec $i_1 < i_2 < i_3$; si un tel croisement existe, on peut construire un triplet w_i w_{i+1} $w_{i'}$, comportant donc deux éléments consécutifs dont la distance (au sens de δ_H) est strictement supérieure à $\delta_H(w_i, w_{i'})$, (i.e. δ_H et Θ ne sont pas faiblement compatibles) ; en effet, soit i (resp. i') le premier indice supérieur (resp. inférieur) à i_1 (resp. i_3) tel que $w_{i+1} \notin h_1$ (resp. $w_{i'-1} \notin h_1$), voir la figure 14.

On a nécessairement $i \neq i'$ puisque $w_{i_2} \notin h_1$ et $i_1 < i_2 < i_3$.

Soit h_1' le plus bas des paliers contenant w_i et w_{i+1} ; on a donc $h_1 \subset h_1'$ puisque $h_1 \cap h_1' \neq \phi$ (à cause de w_i) et w_{i+1} appartient à h_1' et non à h_1 ; il en résulte que $\delta_H(w_i,w_{i'}) < \delta_H(w_i,w_{i+1})$ puisque H est une hiérarchie indicée ; c'est bien la condition cherchée. Il en résulte qu'une condition suffisante pour que Θ ne donne pas lieu à un croisement est que δ_H et Θ soient faiblement compatibles. ∎

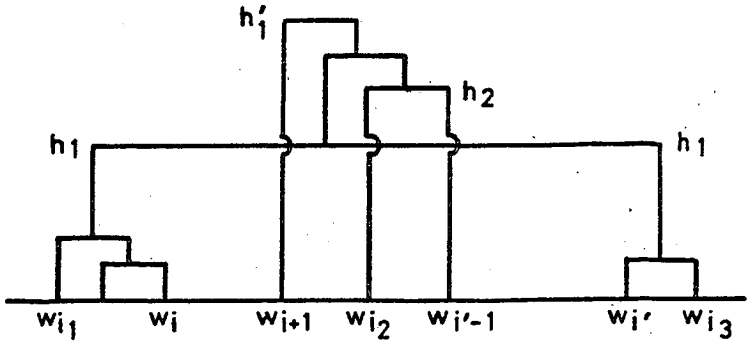


Figure 3

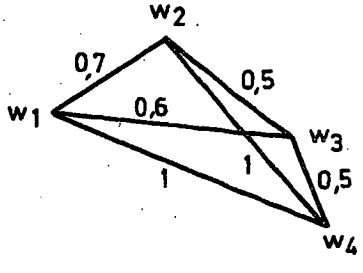
4. - CHAINES DE PLUS COURTES LONGUEUR ET COMPATIBILITE FAIBLE

Un ordre Θ étant donné, on peut lui associer la chaîne hamiltonienne définie sur Ω et dont les arêtes sont formées de deux sommets consécutifs de l'ordre. Un indice de dissimilarité d étant donné, on associe à chaque arête $w_i\ w_{i+1}$ le poids $d(w_i,w_{i+1})$. Une chaîne ainsi valuée sera notée $C(d,\Theta)$. Une chaîne $C(d,\Theta)$ est de plus courte longueur si la somme des poids des arêtes (ainsi définis) est minimum.

Le fait que d et Θ soient faiblement compatibles n'est ni nécessaire, ni suffisant pour que la chaîne $C(d,\Theta)$ soit de plus courte longueur. Ceci est prouvé par les deux exemples suivants. Le premier exemple montre que la condition n'est pas nécessaire et le second qu'elle n'est pas suffisante.

Exemple 1

Considérons la matrice de dissimilarité indiquée figure 4 ; la chaîne $C(d,\Theta)$ induite par l'ordre Θ : $w_1\ w_2\ w_3\ w_4$ est de plus courte longueur bien que d et Θ ne soient pas faiblement compatibles puisque : $d(w_1,w_2) > d(w_1,w_3)$.

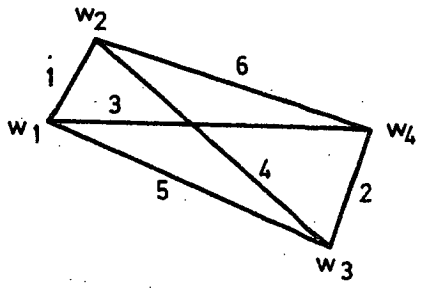


	w_1	w_2	w_3	w_4
w_1	0	0,7	0,6	1
w_2		0	0,5	1
w_3			0	0,5
w_4				0

Figure 4

Exemple* 2

Considérons la matrice de dissimilarité indiquée figure 5 ; la chaîne $C(d,\Theta)$ définie par l'ordre Θ : $w_1\ w_2\ w_3\ w_4$ (de longueur égale à 7) est plus longue que la chaîne $C(d,\Theta')$ définie par Θ' : $w_2\ w_1\ w_4\ w_3$ (de longueur égale à 6) bien que d et Θ soient compatibles :



	w_1	w_2	w_3	w_5
w_1	0	1	5	3
w_2		0	4	6
w_3			0	2
w_4				0

Figure 5

5. - LA SEMI-COMPATIBILITE

5.1. - Définition de la semi-compatibilité

La semi-compatibilité entre un ordre Θ et une distance d est une condition plus restrictive que la compatibilité faible car elle assure à la chaîne $C(d,\Theta)$ d'être de plus courte longueur.

Définition

Nous dirons que d et Θ sont semi-compatibles si et seulement si tout quadruplet $w_{i_1}\ w_{i_2}\ w_{i_3}\ w_{i_4}$ avec $i_1 \leq i_2 < i_3 \leq i_4$ et $i_3 = i_2 + 1$ est tel que $d(w_{i_2}, w_{i_3}) \leq d(w_{i_1}, w_{i_4})$.

Il résulte de cette définition que si d et Θ sont semi-compatibles alors d et Θ sont faiblement compatibles, l'inverse n'étant pas nécessairement vrai.

---

* Ce contre-exemple m'a été signalé par B. Montjardet que je tiens à remercier ici.

5.2. - Semi-compatibilité et chaîne de plus courte longueur

Proposition 2

Si d et Θ sont semi-compatibles, alors la chaîne $C(d,\Theta)$ est de plus courte longueur.

Démonstration

Soit C une chaîne induite par l'ordre Θ noté $w_1,\ldots,w_n$ tel que d et Θ soient semi-compatibles ; soit C^* une chaîne de plus courte longueur au sens de d ; nous allons montrer que l'on peut construire à partir de C^* une suite d'arbres $A_1,\ldots,A_n$ de longueur inférieure ou égale à la longueur de C^* et telle que A_n soit identique à C. On part de $A_1 \equiv C^*$.

L'arbre A_2 s'obtient de la façon suivante : on considère dans C^* le premier sommet w_1 de la chaîne C ; on a $d(w_1,w_2) \leq d(w_1,w_i)\ \forall i \neq 1$ car, par hypothèse, Θ et d sont semi-compatibles. Soit $w_1\ w_\ell$ la première arête de la chaîne qui relie w_1 à w_2 dans C^*. Si $w_\ell \neq w_2$, on supprime cette arête; on obtient ainsi deux parties connexesde C^*; en reliant ces parties par l'arête $w_1\ w_2$, on obtient l'arbre A_1 qui est par construction de longueur inférieure à C^*. Si $w_\ell = w_2$, on pose $A_2 = A_1$.

On construit l'arbre A_{i+1} à partir de l'arbre A_i de la même façon ; si $w_i\ w_{i+1}$ est dans A_i, on pose $A_{i+1} \equiv A_i$, sinon, dans la chaîne issue de A_i qui relie w_i à w_{i+1}, on considère les arêtes qui se succèdent à partir de w_i et on supprime la première arête notée $w_\ell\ w_{\ell'}$ qui n'est pas dans C ; on obtient deux parties connexes de A_i que l'on relie par l'arête $w_i\ w_{i+1}$ pour former l'arbre A_{i+1} ; cet arbre est de longueur inférieure à A_i car on a $\ell \leq i < i+1 < \ell'$ d'où $d(w_i,w_{i+1}) \leq d(w_\ell,w_{\ell'})$ puisque d et Θ sont semi-compatibles. A l'étape n, on obtient l'arbre A_n qui contient, par construction, toutes les arêtes de C, c'est donc une chaîne identique à C.∎

6. - CHAINE DE PLUS COURTE LONGUEUR DANS LE CAS D'UNE ULTRAMETRIQUE

6.1. - Chaîne de plus courte longueur et compatibilité faible

Proposition 3

Si δ est une ultramétrique et si $C(\delta,\Theta)$ est une chaîne de plus courte longueur, alors δ et Θ sont faiblement compatibles.

Démonstration

Nous allons montrer que si δ et Θ ne sont pas faiblement compatibles, alors on peut construire une chaîne C' plus courte que la chaîne $C(\delta,\Theta)$, notée C.

En effet, si δ et Θ ne sont pas faiblement compatibles, il existe, par définition, un triplet $w_{i_1}\ w_{i_2}\ w_{i_3}$ ayant 2 sommets consécutifs tel que $\delta(w_{i_1}, w_{i_3}) < \delta(w_{i_1}, w_{i_2})$ où $w_{i_1}\ w_{i_2}$ sont les 2 sommets consécutifs. Considérons la chaîne qui relie w_{i_1} à w_{i_3} dans C ; en remplaçant $w_{i_1}\ w_{i_2}$ par $w_{i_1}\ w_{i_3}$, on obtient un arbre C_1 de longueur plus courte que C (voir figure 6)

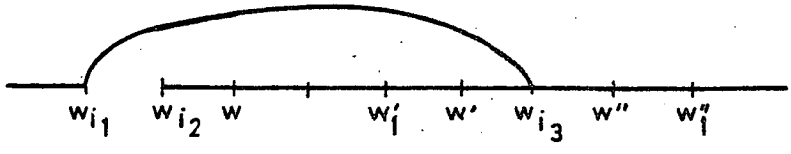
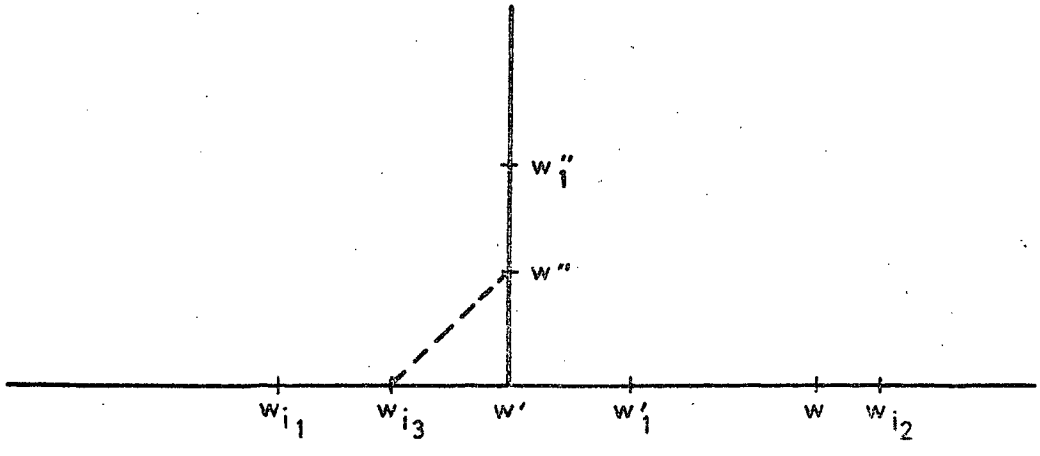


Figure 6

Si C_1 n'est pas une chaîne, il y a deux arêtes de C_1 qui contiennent w_{i_3} ; woit $w_{i_3}\ w'$ et $w_{i_3}\ w''$ ces deux arêtes (voir figure 6) ; le triangle $w'\ w_{i_3}\ w''$ est isocèle avec la base plus petite que les côtés car δ est une ultramétrique ; en remplaçant la plus grande des arêtes $w_{i_3}\ w'$, $w_{i_3}\ w''$, on obtient un nouvel arbre C_2 qui est de longueur inférieure (si $w''\ w''$ est la base du triangle $w'\ w_{i_3}\ w''$) ou égale à C_1 ; si c'est l'arête $w_{i_3}\ w''$ qui a été supprimée, on se trouve dans le cas de la figure 7 :



On peut recommencer le même raisonnement avec le triangle $w'\ w'_1\ w''$ (voir figure 7) en remplaçant soit $w'\ w'_1$ soit $w'\ w''$ par $w''\ w'_1$, on peut obtenir de même un arbre C_3 de longueur inférieure ou égale à C_2 ; on peut recommencer ainsi de suite en écartant de plus en plus de w_{i_3} le sommet d'où partent trois arêtes, jusqu'à obtenir une chaîne qui est strictement plus courte que C ; ce qui est contraire à l'hypothèse ; donc pour que la chaîne C soit de plus courte longueur, il est nécessaire que δ et Θ soient faiblement compatibles. ∎

6.2. - Chaîne de plus courte longueur et croisements

Proposition 4

Si δ_H est l'ultramétrique induite par une hiérarchie indicée H et si la chaîne $C(\delta_H,\Theta)$ est de plus courte longueur alors l'ordre Θ ne donne pas lieu à un croisement pour la hiérarchie H.

Démonstration

D'après la proposition 3, nous savons que si la chaîne $C(\delta_H,\Theta)$ est de plus courte longueur, alors δ_H et Θ sont faiblement compatibles, il en résulte, d'après la proposition 1, que l'ordre Θ ne donne pas lieu à un croisement pour la hiérarchie H. ∎

Remarque : Nous montrerons, plus loin (proposition 10), que la condition donnée dans la proposition 4 est aussi nécessaire afin que l'ordre Θ ne donne pas lieu à un croisement pour la hiérarchie H.

7. - COMPATIBILITE ENTRE UN ORDRE ET UN INDICE DE DISSIMILARITE

Soit d un indice de dissimilarité sur l'ensemble des objets Ω et Θ un ordre sur Ω. On pose $d(w_{\ell_i}, w_{\ell_j}) = d_{ij}$.

Définition (Brossier 1980).

d et Θ sont dits compatibles si et seulement si

$$(1) \qquad w_{\ell_i}\ \Theta\ w_{\ell_j}\ \Theta\ w_{\ell_k} \iff \{d_{ij} \leq d_{ik} \text{ et } d_{jk} \leq d_{ik}\}$$

Il résulte facilement de cette définition que si une distance et un ordre sont compatibles, ils sont faiblement compatibles ; en effet, la condition (1) est à fortiori satisfaite si j = i+1 ou k = j+1. Par contre, si d et Θ sont faiblement compatibles, ils ne sont pas nécessairement compatibles car la condition (1) peut ne pas être satisfaite pour les triplets ne comportant pas de sommets consécutifs. Nous verrons que la semi-compatibilité est une notion intermédiaire entre la compatibilité et la compatibilité faible.Nous verrons également que si d est une ultramétrique, alors les notions de compatibilité, semi-compatibilité et compatibilité faible sont équivalentes.

8. - ASPECTS MATRICIELS : MATRICES DE ROBINSON ET MATRICES SDR ET SDD

8.1. - Matrices de Robinson

Soit D la matrice de dissimilarité associée à d, autrement dit :

$$D = \{d_{ij}\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

La matrice D étant symétrique, on peut définir la matrice de Robinson et les matrices SDR et SDD en ne considérant que la partie triangulaire supérieure de D.

Définition d'une matrice de Robinson (voir, par exemple, Kendall (1969), Hubert (1974))

Une matrice est dite de Robinson si et seulement si les termes des lignes et des colonnes sont croissants à partir de chaque terme de la diagonale (voir figure 8).

0	2	4	6
2	0	4	5
4	4	0	1
6	5	1	0

Matrice de Robinson

0	2	6	5
2	0	4	7
6	4	0	1
5	7	1	0

Matrice SDR

0	2	5	3
2	0	4	6
5	4	0	1
4	6	1	0

Matrice SDD

Figure 8

8.2. - Matrices à sur-diagonale "rectangle"

Considérons la matrice triangulaire supérieure déduite de D ; la diagonale de D étant exclue, la sur-diagonale est la plus grande diagonale de cette matrice. Par exemple, dans la matrice de Robinson, indiquée figure 8, la sur-diagonale est : 2 4 1.

A chaque terme de la sur-diagonale, on peut associer un rectangle dont les côtés sont formés de la ligne et de la colonne contenues dans la matrice triangulaire supérieure et issues de ce terme.

Par exemple, dans la matrice de Robinson, indiquée figure 8, le rectangle (qui est dans ce cas un carré) issu du terme 4 de la sur-diagonale est $\begin{matrix} 4 & 6 \\ 4 & 5 \end{matrix}$.

Définition d'une matrice SDR

Une matrice est dite SDR (sur-diagonale "rectangle") si chaque terme de la sur-diagonale est inférieure aux termes du rectangle qui lui est associé (voir un exemple figure 8).

8.3. - Matrices à sur-diagonale dominée

Définition d'une matrice SDD

Une matrice est dite SDD (sur-diagonale dominée) si dans la matrice triangulaire supérieure associée à D les termes des lignes et des colonnes sont plus grands que le terme de la sur-diagonale qu'elles contiennent (voir figure 8).

On peut résumer ces trois définitions par le tableau 1 (où les intervalles de variation de i, j et ℓ entre 1 et n se déduisent immédiatement des différentes formules ; ainsi par exemple, pour la condition Robinson lignes, on a $i=1,\ldots,n$ et $j=1,\ldots,n-1$).

pour $i \leq j$

$$\text{Robinson} \iff \begin{cases} \text{lignes } d_{ij} \leq d_{ij+1} & \text{SDR} \iff \{d_{ij+1} \leq d_{i\ell} \quad j+1 \leq \ell\} \\ \text{colonnes } d_{ij} \leq d_{i-1j} & \text{SDD} \iff \begin{cases} \text{lignes } d_{ii+1} \leq d_{ij+1} \\ \text{colonnes } d_{j-1j} \leq d_{i-1j} \end{cases} \end{cases}$$

Tableau 1

Il résulte facilement de ces trois définitions que l'ensemble des matrices de Robinson est inclus dans l'ensemble des matrices SDR qui est lui-même inclusdans l'ensemble des matrices SDD.

9. - COMPATIBILITE ET CHAINE DE LONGUEUR MINIMALE

Dans toute la suite, on note $M(d,\Theta)$, une matrice de dissimilarité dont les colonnes respectent l'ordre Θ défini par $w_{\ell_1}, w_{\ell_2}, \ldots, w_{\ell_n}$. Cette matrice s'écrit donc :

$$M(d,\Theta) = \{d_{ij}\}_{\substack{i=1,\ldots,n \\ j=1,\ldots,n}}$$

où rappelons le $d_{ij} = d(w_{\ell_i}, w_{\ell_j})$.

On peut facilement démontrer l'ensemble des résultats suivants (Hubert (1974), Brossier (1980)) :

- Si une matrice $M(d,\Theta)$ est de Robinson, alors la chaîne $C(d,\Theta)$ est de plus courte longueur. Cependant Θ et d peuvent définir une chaîne de plus courte longueur sans que la matrice $M(d,\Theta)$ soit de Robinson comme le prouve l'exemple 1 donné en 4.2 (figure 4). On peut aussi remarquer que si $M(d,\Theta)$ est de Robinson, alors la chaîne associée est aussi un arbre de longueur minimum (pour le montrer, il suffit d'utiliser l'algorithme de Prim à partir du premier sommet de la chaîne défini par l'ordre Θ).

Proposition 5 (Brossier (1980))

Une condition nécessaire et suffisante pour qu'une matrice M(d,Θ) soit de Robinson et que d et Θ soient compatibles.

Démonstration

L'ordre Θ étant défini par la suite $w_{\ell_1}, w_{\ell_2}, \ldots, w_{\ell_n}$, on a les équivalences :

$$d \text{ et } \Theta \text{ compatibles } \Leftrightarrow \left\{ \begin{array}{ll} d_{ij} \leq d_{ik} & \\ & \text{si } i \leq j \leq k \\ d_{jk} \leq d_{ik} & \end{array} \right\}$$

$$\Leftrightarrow \left\{ \begin{array}{ll} d_{ij} \leq d_{ik} & \\ & \text{si } i \leq j \leq k \\ d_{jk} \leq d_{ik} & \end{array} \right\}$$

$$\Leftrightarrow \left\{ \begin{array}{ll} d_{ij} \leq d_{ij+1} & \\ & \text{si } i \leq j \\ d_{ij} \leq d_{i-1j} & \end{array} \right\}$$

$\Leftrightarrow$ {M(d,Θ) est une matrice de Robinson} ∎

10. - SEMI-COMPATIBILITE MATRICE SDR, CHAINE DE PLUS COURTE LONGUEUR ET ARBRE DE LONGUEUR MINIMUM

Proposition 6

Une condition nécessaire et suffisante pour qu'une matrice M(d,Θ) soit SDR est que d et Θ soient semi-compatibles.

Démonstration

En effet, la définition d'une matrice SDR conduit à l'équivalence donnée dans le tableau 1 qui revient à dire que d et Θ sont semi-compatibles (il suffit de remplacer, dans la définition de la semi-compatibilité i_1 par i, i_2 par j et i_4 par ℓ). ∎

Proposition 7

Si M(d,Θ) est SDR, alors la chaîne induite par Θ est de plus courte longueur au sens de d.

Démonstration

Si M(d,Θ) est SDR,alors d et Θ sont semi-compatibles d'après la proposition 6 et donc la chaîne induite par Θ est de plus courte longueur d'après la proposition 2. ∎

L'idée revient à B. Monjardet (1982) d'avoir vu le lien entre la notion de semi-compatibilité et de chaîne T-minimax de B. Leclerc (1974,1981). D'où la définition de la semi-compatibilité partielle et le lien entre cette notion et les arbres de longueur minimum que nous proposons maintenant.

Définissons d'abord la notion d'élément à gauche et d'élément à droite d'une chaîne $C(d,\Theta')$ où $\Theta' : w_{i_1},\ldots,w_{i_\ell}$ avec $\ell \leq n$ est un ordre partiel sur Ω.

. Un élément est à gauche (resp. à droite de la chaîne $C(d,\Theta')$ si et seulement si sa distance, à tout point w de la chaîne, est supérieure à la distance de deux points consécutifs qui précèdent (resp. suivent) w selon l'ordre Θ', s'il est à l'extérieur de la chaîne. S'il est dans la chaîne, sa distance à tout point w qu'il précède (resp. qu'il suit) selon l'ordre Θ', doit être supérieure à tout couple d'éléments consécutifs intermédiaires entre lui et w.

. Autrement dit, w est un élément à gauche si et seulement si $d(w,w_{i_j}) \geq d(w_{i_{k-1}},w_{i_k})$ pour tout $j = m+1,\ldots,n$, $k = m+1,\ldots,j$ où m est le rang de w dans la chaîne quand w est dans la chaîne et m = 1 sinon.

. De même un élément est à droite de la chaîne $C(d,\Theta')$ si et seulement si $d(w,w_{i_j}) \geq d(w_{i_k},w_{i_{k+1}})$ avec $j = 1,\ldots,m-1$, $k = j,\ldots,m-1$ où m a la même valeur que précédemment.

Définition*

Un indice de dissimilarité d et un ordre Θ' sur une partie de Ω sont partiellement semi-compatibles sur Ω si et seulement si tous les éléments de Ω sont soit à gauche, soit à droite de la chaîne $C(d,\Theta')$ et si la distance entre un point à gauche et un point à droite de cette chaîne est supérieure à la distance entre deux points consécutifs selon l'ordre Θ'.

Il résulte facilement de cette définition que d et Θ' sont semi-compatibles sur la partie de Ω formée des individus qui définissent l'ordre Θ'.

La condition nécessaire (Hu (1960) et Kalaba (1964)) et suffisante (Leclerc (1974)) pour qu'une chaîne soit T-minimax est qu'elle relie deux éléments d'un arbre de longueur minimum. De façon analogue à ce résultat, nous avons la proposition suivante :

Proposition 8 (Hu, Kalaba, Leclerc)

Une c.n.s. pour qu'un ordre Θ' : $w_{i_1} \ldots w_{i_\ell}$ $(\ell \leq l)$ sur une partie de Ω et un indice de dissimilarité d soient partiellement semi-compatibles sur Ω est que la chaîne $C(d,\Theta')$ relie w_{i_1} à w_{i_ℓ} dans un arbre de longueur minimum.

Démonstration

La condition est nécessaire (Hu, Kaleba) ; en effet, considérons une chaîne C dans un arbre de longueur minimum A. Si l'on coupe une arête $w_{i_j}\ w_{i_{j+1}}$ de cette chaîne, on obtient deux parties connexes de l'arbre. En reliant deux éléments w' w" pris dans chacune des parties connexes, on obtient un nouvel arbre A' qui est certainement de longueur supérieure à A. Il en résulte que $d(w',w'') \geq d(w_{i_{j-1}}, w_{i_j})$; pour montrer que w' est un élément à gauche, il faut considérer deux cas :

1) w' est un élément de la chaîne et alors il existe t : $1 \leq t \leq j$ tel que $w' = w_{i_t}$.

2) w' est extérieur à la chaîne.

---

* Une définition équivalente mais plus générale est donnée § 22 p. 53, ainsi que les aspects matriciels.

En coupant l'arbre A suivant les arêtes $w_{i_{k-1}}$ w_{i_k} pour $k = t,t+1,..,j$ avec $t \geq 1$ dans le premier cas et $t = 1$ dans le second, on obtient successivement les inégalités $d(w',w_{i_j}) \geq d(w_{i_{k-1}},w_{i_k})$, en remarquant que w' reste dans la partie connexe de A qui ne contient pas w_{i_j}. Comme on peut recommencer le même raisonnement pour tout j=1,...,n, on voit que w' est bien un élément à gauche de la chaîne. On peut montrer, de façon analogue, que w" est un élément à droite de la chaîne, d'où la condition nécessaire.

Montrons maintenant la condition suffisante ; considérons une chaîne $C(d,\Theta')$: $w_{i_1} \ldots w_{i_\ell}$, $\ell \leq n$. Construisons l'arbre de longueur minimum à l'aide de l'algorithme de Prim en partant du premier élément de cette chaîne. Supposons que les j-1 premiers éléments de la chaîne soient les premiers éléments introduits par l'algorithme de Prim dans l'arbre de longueur minimum. Quand c'est au tour de l'élément w_{i_j} d'entrer dans l'arbre de longueur minimum par l'algorithme de Prim, il est le premier à entrer parmi les éléments w" à droite de la chaîne puisque : $d(w",w_{i_k}) \geq d(w_{i_{j-1}},w_{i_j})$ pour tout k = 1, 2, ..., j-1. Il ne peut se connecter avec un élément w' à gauche qui serait déjà entré puisque $d(w',w_{i_{j+1}}) \geq d(w_{i_j},w_{i_{j+1}})$, il se connecte donc nécessairement à $w_{i_{j-1}}$. Comme on peut recommencer le même raisonnement pour j = 2 à ℓ- 1, on a le résultat. ∎

B. Leclerc(1974) montre que les chaînes T-minimax hamiltoniennes caractérisent les arbres de longueur minimum d'où par analogie le résultat suivant qui se déduit immédiatement de la proposition précédente et qui avait déjà été énoncé par Rosensthiel.

Corollaire (Rosensthiel (1967))

Une c.n.s pour qu'une chaîne $C(d,\Theta)$ où Θ est un ordre total soit un arbre de longueur minimum est que d et Θ soient semi-compatibles.

Démonstration

Il suffit d'utiliser la proposition 8 en remarquant que si Θ' est un ordre total, les notions de semi-compatibilité partielle sur Ω et de semi-compatibilité sont équivalentes, d'une part, et que $C(d,\Theta')$ est un arbre, d'autre part. ∎

Gau et Fichet ont signalé une démonstration directe de ce corollaire ; en voici une :

Démonstration

Soit Θ l'ordre défini par la suite $w_{i_1}, \ldots, w_{i_n}$. La condition est nécessaire. En effet, si la matrice $M(d,\Theta)$ n'était pas SDR, il existerait au moins un couple (w_{i_ℓ}, w_{i_j}) avec $j > \ell$ tel que $d(w_{i_\ell}, w_{i_j}) < d(w_{i_{j-1}}, w_{i_j})$; pour que la chaîne $C(d,\Theta)$ soit un arbre de plus courte longueur, il faut qu'en partant de w_{i_1} l'algorithme de Prim construise la chaîne induite par Θ ; or, même si à l'étape $j-1$, la partie de l'arbre déjà construite est la chaîne $w_{i_1}, w_{i_2}, \ldots, w_{i_{j-1}}$ et l'élément le plus proche des sommets déjà atteints est w_{i_j}, la nouvelle arête n'est certainement pas $w_{i_{j-1}}\ w_{i_j}$ puisque l'arête $w_{i_\ell}\ w_{i_j}$ est strictement plus courte ; il en résulte que $C(d,\Theta)$ n'est pas un arbre de plus courte longueur.

Montrons maintenant que la condition est suffisante ; supposons qu'à l'étape $j-1$ l'algorithme de Prim donne la chaîne $w_{i_1}, w_{i_2}, \ldots, w_{i_{j-1}}$; il suffit de montrer que si $M(d,\Theta)$ est SDR, alors l'arête suivante sera $w_{i_{j-1}}\ w_{i_j}$; en effet, si $M(d,\Theta)$ est SDR, un sommet qui est le plus proche de $w_{i_1}, w_{i_2}, \ldots, w_{i_{j-1}}$ est forcément w_{i_j} puisque, par définition d'une matrice SDR, on a $d(w_{i_j}, w_{i_{j-1}}) \leq d(w_{i_\ell}, w_{i_q})\ \forall\, \ell \leq j$ et $j - 1 \leq q$. Pour la même raison, $w_{i_j}\ w_{i_{j-1}}$ est une arête de raccordement de plus courte longueur. ∎

Proposition 9

Si la chaîne $C(d,\Theta)$ est un arbre de plus courte longueur, alors l'ordre Θ est sans croisement pour une hiérarchie du saut minimum.

Démonstration

Etant donné un arbre de longueur minimum, on peut lui associer une hiérarchie du saut minimum (celle qui est associée à l'indice d'agrégation du lien minimum) telle qu'à chaque palier de cette hiérarchie il corresponde une partie connexe de cet arbre (c'est un résultat que l'on trouvera par exemple dans [ ]) ; l'ordre Θ ne donne pas lieu à un croisement pour cette hiérarchie ; en effet, s'il y avait un croisement, il existerait, par définition, un palier h tel que : $w_{i_\ell} \in h$, $w_{i_q} \in h$, $w_{i_p} \notin h$ avec $\ell < p < q$. La partie de l'arbre de longueur minimum, définie par la chaîne $C(d,\Theta)$, associée au palier h_1, ne serait donc pas connexe. ■

Remarque : Il résulte des propositions 8 et 9 que si $M(d,\Theta)$ est SDR, alors l'ordre Θ est sans croisement pour la hiérarchie du saut minimum construite à l'aide de l'indice de dissimilarité d ; par contre, quand un ordre Θ est sans croisement pour une hiérarchie du saut minimum construite à l'aide d'une mesure de dissimilarité d, la matrice $M(d,\Theta)$ n'est pas nécessairement SDR, comme le montre l'exemple suivant.

Exemple

On a : $\Omega = \{w_1, w_2, w_3\}$; $\Theta : w_1\ w_2\ w_3$; la matrice $M(d,\Theta)$ est donnée figure 9 et la hiérarchie du saut minimum (elle est unique ici) qui lui est associée, figure 10.

	w_1	w_2	w_3
w_1	0	1	2
w_2		0	3
w_3			0

Figure 9

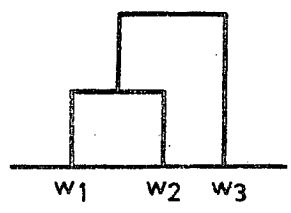


Figure 10

On voit que l'ordre Θ est sans croisement pour cette hiérarchie bien que $M(d,\Theta)$ ne soit pas SDR. On peut remarquer aussi que l'ordre Θ' : $w_3\ w_1\ w_2$ qui rend la matrice $M(d,\Theta')$ SDR est bien sans croisement pour cette hiérarchie.

II.- COMPATIBILITE FAIBLE ET MATRICE SDD

On a un résultat analogue au précédent en utilisant les matrices SDD.

Proposition 10

Une condition nécessaire et suffisante pour qu'une matrice $M(d,\Theta)$ soit SDD est que d et Θ soient faiblement compatibles.

Démonstration

Θ étant défini par l'ordre $w_{\ell_1},\ldots,w_{\ell_n}$, on a la suite d'équivalences suivantes :

{d et Θ faiblement compatibles}

$$\Leftrightarrow \left\{ \forall (w_{\ell_i}, w_{\ell_j}, w_{\ell_k}) \in \Omega^3 \quad \begin{array}{l} k-1 \geq j=i+1 \Rightarrow d_{ii+1} \leq d_{ik} \\ k-1 = j \geq i+1 \Rightarrow d_{jj+1} \leq d_{ij+1} \end{array} \right\}$$

$$\Leftrightarrow \left\{ \begin{array}{l} d_{i\ i+1} \leq d_{i\ell+1} \text{ si } i \leq \ell \text{ (on pose } \ell = k-1) \\ d_{j\ j+1} \leq d_{ij+1} \text{ si } i \leq j-1 \text{ et si } i=j \text{ donc } i \leq j \end{array} \right\}$$

$\Leftrightarrow$ {M(d,Θ) est une matrice SDD}. ∎

On peut résumer l'ensemble de ces propriétés par le schéma de la figure 11.

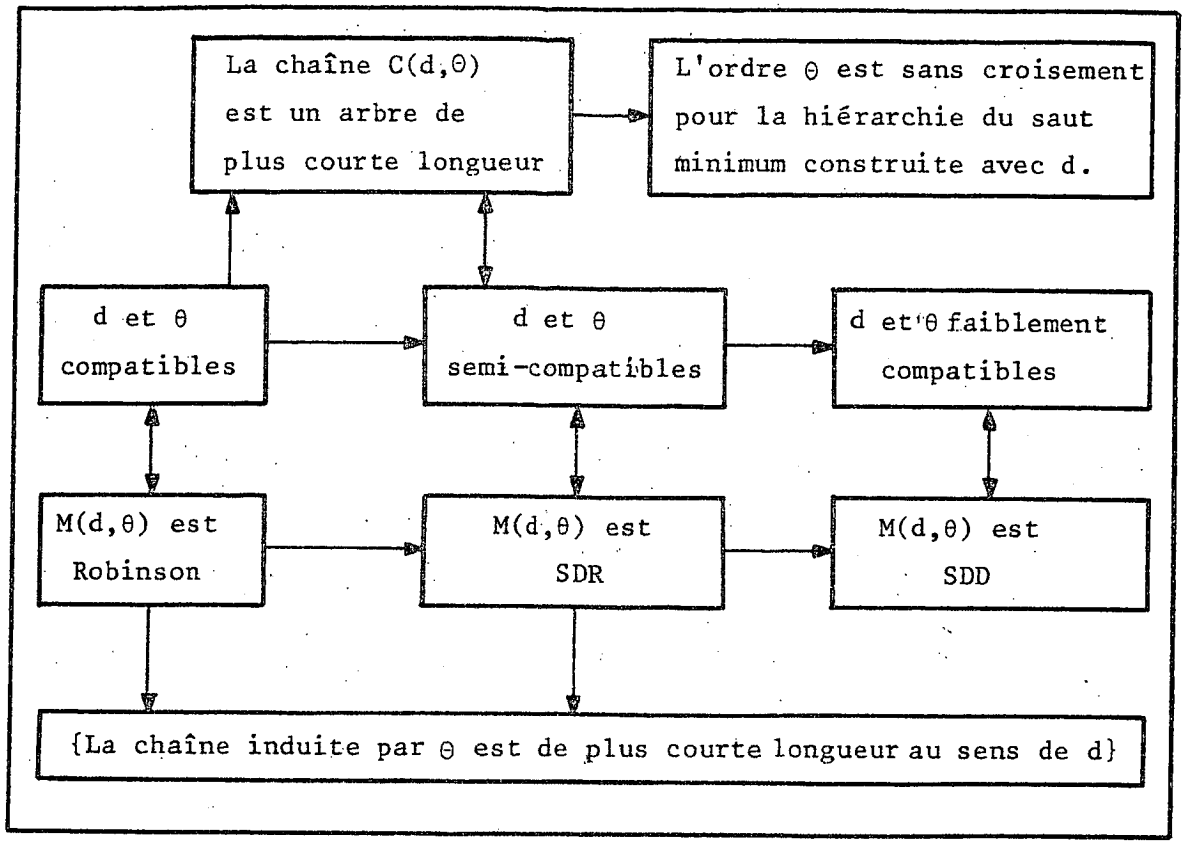


Figure 11

12.- CAS D'UNE ULTRAMETRIQUE

Dans le cas où d est une ultramétrique que nous noterons δ, le schéma de la figure 11 se transforme en plaçant des équivalences là où il y avait des implications. On a en effet, les deux propositions suivantes :

Proposition 11

Si δ est une ultramétrique et si $M(\delta,\Theta)$ est SDD, alors $M(\delta,\Theta)$ est une matrice de Robinson.

Démonstration

Si $M(\delta,\Theta)$ est SDD, on a $\delta_{ij+1} \geq \delta_{jj+1}$ pour $i \leq j$, si de plus δ est une ultramétrique, on a : $\delta_{ij} \leq \text{Max}\ (\delta_{ij+1}, \delta_{jj+1})$; d'où $\{\delta_{ij} \leq \delta_{ij+1}$ pour $i \leq j\}$ qui est la condition "ligne" (voir tableau 1) à satisfaire par $M(\delta,\Theta)$ pour être de Robinson.

De même si $M(\delta,\theta)$ est SDD, on a $\delta_{i-1j} \geq \delta_{j-1j}$, en utilisant de même le fait que δ est une ultramétrique, on a $\delta_{ij} \leq \text{Max}\ (\delta_{i-1j},\delta_{i-1i}) = \delta_{i-1j}$ si $i \leq j$ d'où la condition "colonne" à satisfaire par $M(\delta,\theta)$ pour être de Robinson. ∎

Signalons que I.C. Lerman (1981) a donné aussi une caractérisation d'une matrice $M(\delta,\theta)$.

Proposition 12

Si δ est une ultramétrique et si $C(\delta,\theta)$ est une chaîne de plus courte longueur, alors la matrice $M(\delta,\theta)$ est SDD.

Démonstration

Si la chaîne induite par θ est de plus courte longueur au sens d'une ultramétrique δ, nous savons, d'après la proposition 3 que δ et θ sont faiblement compatibles donc que la matrice $M(\delta,\theta)$ est SDD d'après la proposition 10. ∎

Proposition 13

Une condition nécessaire et suffisante pour qu'un ordre θ soit sans croisement pour une hiérarchie H est que la chaîne induite par θ soit de plus courte longueur au sens de l'ultramétrique δ_H induite par H.

Démonstration

La condition suffisante résulte de la proposition 4. La condition nécessaire est prouvée par la suite des implications suivantes :

Si l'ordre θ est sans croisement pour la hiérarchie H, alors δ_H et θ sont faiblement compatibles d'après la proposition 1 ; la matrice $M(\delta_H,\theta)$ est donc SDD d'après la proposition 10, c'est donc une matrice de Robinson d'après la proposition 11 et donc la chaîne $C(\delta_H,\theta)$ est de plus courte longueur. ∎

Avant de présenter le schéma qui résume l'ensemble des résultats obtenus dans le cas d'une ultramétrique, nous donnons un dernier résultat qui conduit à l'énoncé d'un algorithme très simple pour la construction d'une chaîne de plus courte longueur dans le cas d'une ultramétrique.

Nous dirons qu'une matrice M(d,Θ) est SDDL (resp. SDDC) si dans la matrice triangulaire supérieure associée à M(d,Θ) chaque terme de la sur-diagonale est inférieur à tous les termes de la ligne (resp. colonne) qui le contient (voir figure 12).

0	2	3	5
2	0	4	6
3	4	0	2
5	6	2	0

Matrice SDDL

0	2	5	4
2	0	3	2
5	3	0	1
4	2	1	0

Matrice SDDC

Figure 12

Plus précisément une matrice est SDDL (resp. SDDC) si elle satisfait à la condition "ligne" (resp."colonne") donnée dans le tableau 1 ; autrement dit, M(d,Θ) est SDDL si $d_{ii+1} \leq d_{ij+1}$ pour $i \leq j$, elle est SDDC si $d_{j-1j} \leq d_{i-1j}$ pour $i \leq j$. A l'aide de ces définitions, on peut énoncer le résultat suivant :

Proposition 14

Si δ est une ultramétrique et si une matrice est SDDC ou SDDL, alors elle est SDD.

Démonstration

Si δ est une ultramétrique on a $\delta_{i+1j+1} \leq \text{Max}(\delta_{ii+1}, \delta_{ij+1})$; si de plus M(δ,Θ) est SDDL, on a : $\delta_{ii+1} \leq \delta_{ij+1}$ pour $i \leq j$; d'où $\delta_{i+1j+1} \leq \delta_{ij+1}$; en reprenant le même raisonnement, un nombre fini de fois (j+1-i fois), on obtient la suite d'inégalités :

$$\delta_{ij+1} \geq \delta_{i+1j+1} \geq \delta_{i+2j+1} > \cdots \geq \delta_{j-1j+1} \geq \delta_{jj+1}$$

d'où $\{\delta_{ij+1} \leq \delta_{jj+1}$ pour $i \leq j\}$ qui s'écrit aussi en faisant la transformation $i \to i-1$ et $j \to j-1$ $\{\delta_{j-1j} \leq \delta_{i-1j}$ pour $i \leq j\}$ qui prouve que la matrice M(δ,Θ) est SDDC.

On démontre de façon tout à fait analogue qu'une matrice SDDC est SDDL ; on a :

$$\delta_{ii+1} \leq \delta_{ii+2} \leq \ldots \leq \delta_{ij-1} \leq \delta_{ij} \leq \text{Max}(\delta_{jj+1}, \delta_{ij+1}) = \delta_{ij+1}$$

d'où $\{\delta_{ii+1} \leq \delta_{ij+1}$ pour $i \leq j\}$ qui prouve que $M(\delta,\theta)$ est SDDL.

Ainsi quand δ est une ultramétrique, une matrice $M(\delta,\theta)$ qui est SDDL (resp. SDDC) étant aussi SDDC (resp. SDDL) elle est SDD. ■

On peut résumer l'ensemble des résultats obtenus dans le cas d'une ultramétrique à l'aide du schéma de la figure 13 où l'on note H la hiérarchie indicée associée à δ.

Pour retrouver toutes les cases du tableau de la figure 11, on remarque facilement que, par construction même, H_δ est la hiérarchie du saut minimum construite avec l'indice δ.

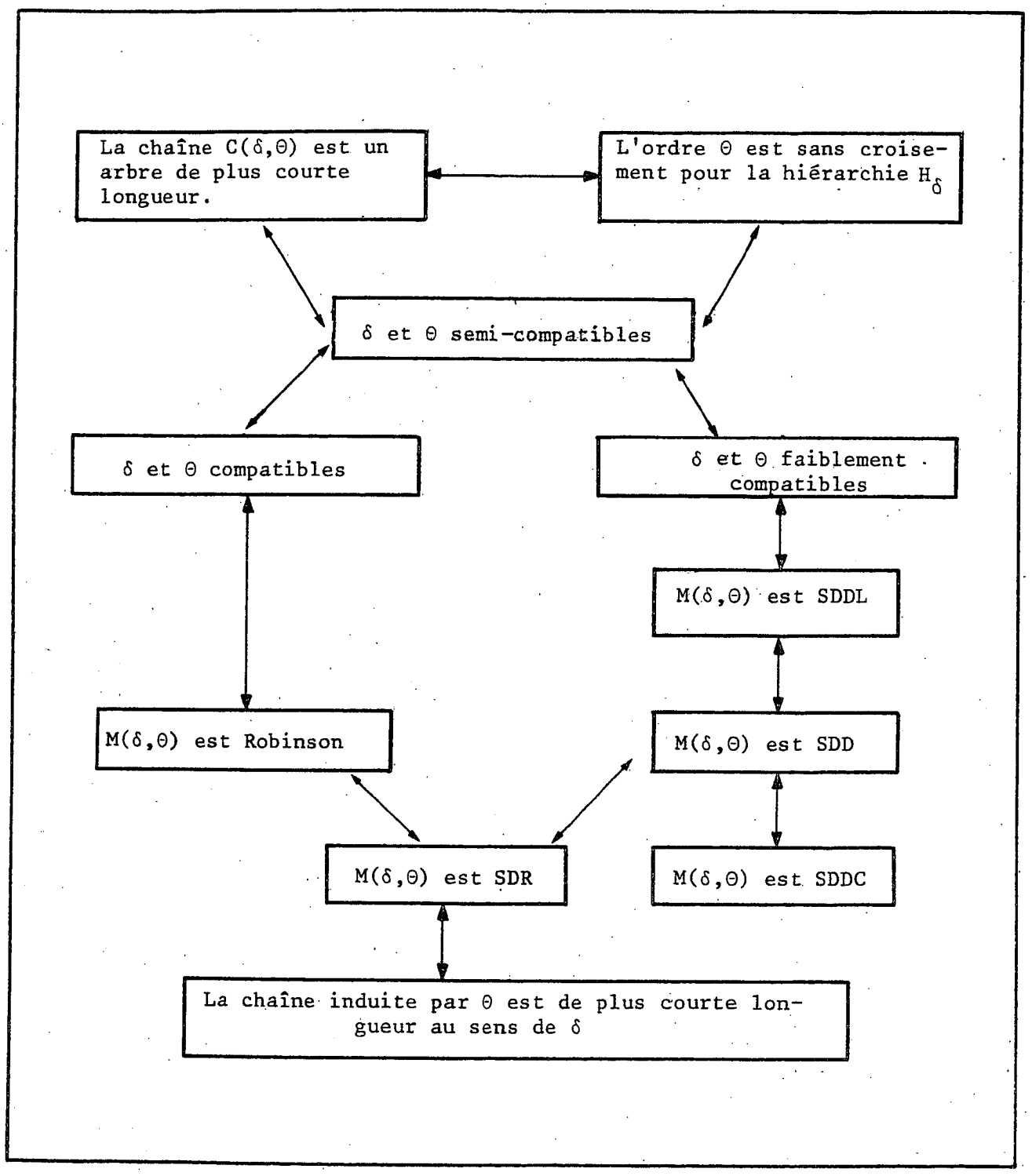


Figure 13 - Cas d'une ultramétirque δ induite par une hiérarchie indicée H_δ

13.- CALCUL RAPIDE D'UNE CHAINE DE PLUS COURTE LONGUEUR DANS LE CAS D'UNE ULTRAMETRIQUE

Etant donnée une ultramétrique δ, l'algorithme (voir figure 14, l'organigramme structuré) consiste à construire une suite $w_{i_1}\ w_{i_2}\ \ldots\ w_{i_n}$ d'éléments de Ω de la façon suivante : w_{i_1} est choisi au hasard, à l'étape ℓ, $w_{i_{\ell+1}}$ est le plus proche de w_{i_ℓ} au sens de δ parmi les éléments qui ne sont pas déjà dans la suite. Autrement dit :

$$\delta(w_{i_{\ell+1}}, w_{i_\ell}) = \text{Min}\{\delta(w_{i_j}, w_{i_\ell}) \ / \ j > \ell\}$$

A chaque étape, dans le cas où le minimum n'est pas unique $w_{i_{\ell+1}}$ est choisi au hasard parmi les termes qui réalisent le minimum. Remarquons que cet algorithme nécessite $n-1 + n-2 + \ldots + 1$ comparaisons $= \frac{(n-1)n}{2} \simeq \frac{n^2}{2}$ comparaisons.

Proposition 15

Cet algorithme construit une chaîne de plus courte longueur.

Démonstration

Soit Θ l'ordre défini par la suite construite par cet algorithme ; il est facile de voir que la matrice $M(\delta,\Theta)$ est SDDL et donc que la chaîne $C(\delta,\Theta)$ est de plus courte longueur. ∎

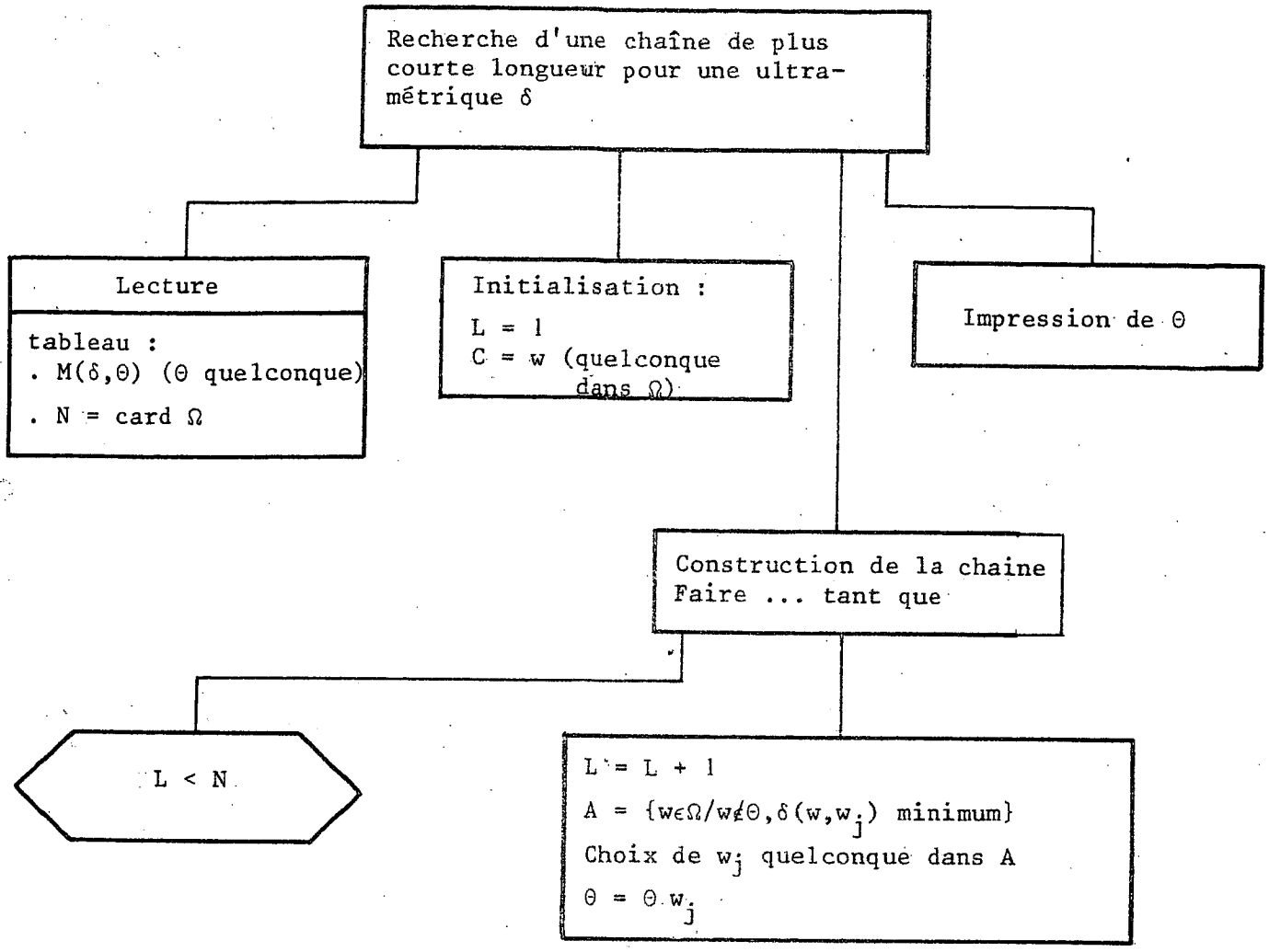


Figure 14

14.- COMPARAISON VISUELLE DE HIERARCHIE

14.1.- Le problème

Dans ce paragraphe, nous nous intéressons aux aspects visuels d'une hiérarchie et notre problème est le suivant : comment visualiser, de façon proche, deux hiérarchies ?

Plus précisément, notons E_i l'ensemble des ordres qui ne donnent pas lieu à un croisement pour la hiérarchie H_i, (rappelons (voir 3.1) que si H_i est une hiérarchie binaire card $E_i = 2^{n-1}$ où n = card Ω) ; le problème consiste alors à chercher $\Theta_1 \in E_1$ et $\Theta_2 \in E_2$ tels que :

i) $\Theta_1 \equiv \Theta_2$ si $E_1 \cap E_2 \neq \phi$

ii) les séquences communes à Θ_1 et Θ_2 contiennent le plus grand nombre d'éléments si $E_1 \cap E_2 = \phi$.

Remarquons que seuls les trois cas suivants peuvent se produire :

$E_1 \equiv E_2$, $E_1 \cap E_2 \neq \phi$ et $E_1 \cap E_2 = \phi$.

Exemple

Le cas $E_1 \equiv E_2$ est illustré par la figure 15 :

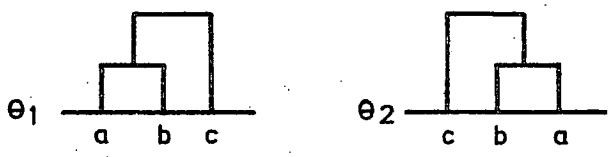


Figure 15

$E_1 \equiv \{(abc), (bac), (cab), (cba)\} \equiv E_2$;

dans ce cas tout élément de E_i (i=1,2) est solution du problème.

Le cas $E_1 \cap E_2 \neq \phi$ est illustré par la figure 16 :

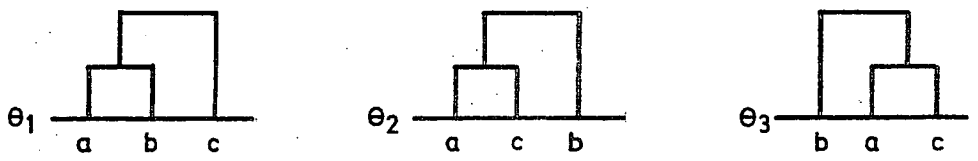


Figure 16

On a :

$E_1 \equiv \{(abc), (bac), (cab), (cba)\}$

$E_2 \equiv \{(acb), (cab), (bac), (bca)\}$

donc :

$$E_1 \cap E_2 = \{(bac), (cab)\}$$

Une solution au problème est donc : $\Theta_3 = (bac)$.

Enfin le cas $E_1 \cap E_2 = \phi$ est donné figure 17 :

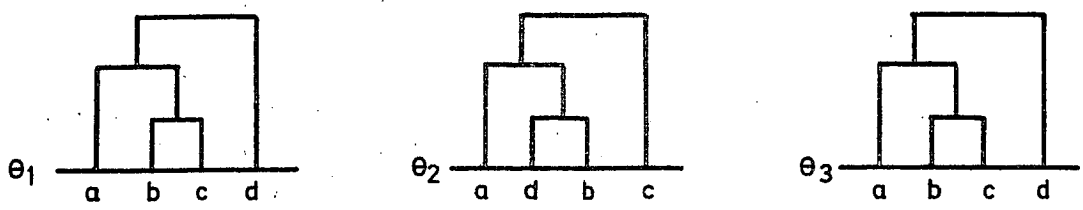


Figure 17

$$E_1 = \{(abcd), (acbd), (bcad), (cbad), (dabc), (dacb), (dbca), (dcba)\}$$

$$E_2 = \{(adbc), (abdc), (dbac), (bdac), (cadb), (cabd), (cdab), (cbda)\}$$

Donc :

$$E_1 \cap E_2 = \phi$$

Un ordre donnant les grandes séquences communes est $\Theta_3 = (abc)$ (il n'y en a qu'une c'est (ab)).

Avant de donner un algorithme permettant de donner une solution à ces problèmes, remarquons que l'on peut savoir, par un calcul simple, si E_1 n'est pas identique à E_2 grâce à la proposition suivante :

Proposition 16

Si δ_i est l'ultramétrique induite par la hiérarchie indicée H_i, $\Theta_i \in E_i$ et $L(\delta_i,\Theta_j)$ est la longueur de la chaîne $C(\delta_i,\Theta_j)$ alors la condition $L(\delta_i,\Theta_1) - L(\delta_i,\Theta_2) \neq 0$ implique $E_1 \neq E_2$.

Démonstration

Si $E_1 \equiv E_2$, alors Θ_1 et Θ_2 ne donnent pas lieu à un croisement pour les hiérarchies H_1 et H_2. Il en résulte (voir figure 13) que $C(\delta_i,\Theta_1)$ et $C(\delta_i,\Theta_2)$ sont de plus courte longueur (pour i = 1,2) d'où :

$$L(\delta_i,\Theta_1) - L(\delta_i,\Theta_2) = 0$$

donc si :

$$L(\delta_i,\Theta_1) - L(\delta_i,\Theta_2) \neq 0$$

on a nécessairement E_1 non identique à E_2. ∎

14.2.- Un algorithme donnant une chaîne sans croisement pour deux hiérarchies

Le but de cet algorithme est de fournir, à partir des ultramétriques δ_1 et δ_2 induites par deux hiérarchies indicées H_1 et H_2, un ordre Θ (s'il existe) qui soit sans croisement pour ces deux hiérarchies. Il consiste à calculer une chaîne de plus courte longueur simultanément au sens de δ_1 et δ_2 en utilisant l'algorithme du § 13. On part d'un même élément w_{i_1} de Ω qui constitue le premier élément de la chaîne. On considère les éléments qui sont à distance minimum de w_{i_1} au sens de δ_1 (ils sont notés $A^1_{i_0}$) et au sens de δ_2 (ils sont notés $A^2_{i_0}$). Si $A^1_{i_0} \cap A^2_{i_0}$ n'est pas vide, on choisit un élément w_{i_2} dans cette intersection, il constitue le second élément de la chaîne. On recommence le procédé jusqu'à ce que l'une des deux possibilités suivantes soient atteintes :

i) $A^1_{i_\ell} \cap A^2_{i_\ell}$ est vide, alors on remet en question les choix faits dans les intersections précédentes et on poursuit l'algorithme avec les nouveaux choix.

ii) la chaîne contient tous les éléments de Ω.

Plus précisément, l'algorithme peut être décrit de la façon suivante :

Au départ, on pose $\ell = 1$, $T_1 = \phi$, $A^1_{i_0} = A^2_{i_0} = \Omega$:

1) Choisir w_{i_ℓ} dans $A^1_{i_{\ell-1}} \cap A^2_{i_{\ell-1}}$ avec $w_{i_\ell} \notin T_\ell$; mettre w_{i_ℓ} dans T_ℓ.

2) On note $A^j_{i_\ell}$ l'ensemble des éléments de Ω, distincts de w_{i_q} pour $q = 1,\ldots,\ell$, qui sont à distance minimum de w_{i_ℓ} au sens de δ_j.

Si $A^1_{i_\ell} \cap A^2_{i_\ell} \neq \phi$ on fait $\ell = \ell + 1$ et on retourne en 1) tant que $\ell < n$.

3) Si $A^1_{i_\ell} \cap A^2_{i_\ell} = \phi$, retour à 1) tant que $T_\ell \neq \text{card } A^1_{i_{\ell-1}} \cap A^2_{i_{\ell-1}}$.

4) Si $T_\ell = \text{card } A^1_{i_{\ell-1}} \cap A^2_{i_{\ell-1}}$, faire $T_\ell = \phi$ et $\ell = \ell - 1$ puis retour à 1).

5) On arrête le processus avec l'ordre $\Theta : w_{i_1} \ldots w_{i_\ell}$ si $T_1 = \Omega$ ou si $\ell = n$.

Afin d'énoncer les deux propriétés essentielles concernant l'ordre ainsi obtenu, on est amené à définir la notion de croisement pour un ordre partiel.

Définition

On dit qu'un ordre $\Theta : w_{i_1} \ldots w_{i_\ell}$ avec $\ell \leq n$ donne lieu à un croisement pour la hiérarchie H si et seulement si il existe $h \in H$ contenant deux éléments w_{i_k}, et w_{i_n} de Ω tels qu'il existe $w_{i_m} \notin h$ et $k < m < n \leq \ell$

Proposition 17

1) L'ordre $\Theta : w_{i_1} \ldots w_{i_\ell}$ avec $\ell \leq n$ donné par l'algorithme est sans croisement pour les deux hiérarchies.

2) S'il existe un ordre total (i.e. $\ell = n$) sans croisement pour les deux hiérarchies, alors l'algorithme donne un ordre total qui est également sans croisement pour les deux hiérarchies.

Démonstration

Pour prouver 1), considérons l'ordre $\Theta : w_{i_1} \ldots w_{i_\ell}$ obtenu par l'algorithme ; il a été construit par l'algorithme du § 13 en utilisant les ultramétriques δ_1 et δ_2 ; Si $\ell < n$, on peut poursuivre à partir de w_{i_ℓ} et avec l'ultramétrique δ_1 (resp. δ_2), l'utilisation de cet algorithme,

on obtient ainsi l'ordre Θ_1 (resp. Θ_2) ; ces ordres sont donc sans croisement pour H_1 et H_2 d'après la proposition 15, leur partie commune $w_{i_1} \ldots w_{i_\ell}$ est donc nécessairement sans croisement pour H_1 et H_2. Si $\ell = n$, l'ordre Θ a été construit par l'agorithme du § 13 avec δ_1 et δ_2, il est donc sans croisement pour H_1 et H_2.

Pour prouver 2), considérons la suite à partir de Ω: $B_1 \ldots B_n$ où B_ℓ est l'ensemble des éléments de Ω qui se trouvent à la ième position dans au moins un ordre sans croisement pour les deux hiérarchies. Autrement dit, si S est l'ensemble des ordres sans croisement, on a :

$$B_\ell = \{w_{i_\ell} \in \Omega \;/\; \Theta : w_{i_1} \ldots w_{i_\ell} \ldots w_{i_n},\ \Theta \in S\}.$$

S'il existe un ordre Θ^* : $w^*_{i_1} \ldots w^*_{i_n}$ qui soit sans croisement pour les deux hiérarchies, aucune des parties B_ℓ pour $\ell = 1,\ldots,n$ n'est vide.

Si le premier élément w_{i_1} de l'ordre Θ construit par l'algorithme est en dehors de B_1, alors Θ ne contient pas les n éléments de Ω car s'il les contenait, d'après la propriété 1), Θ serait un ordre sans croisement pour les deux hiérarchies et par définition w_{i_1} appartiendrait à B_1. L'algorithme remet donc en question le choix de w_{i_1} jusqu'à ce que $w_{i_1} \in B_1$.

Supposons que l'élément w_{i_ℓ} de rang ℓ pour l'ordre Θ construit par l'algorithme soit dans B_ℓ, alors $A^1_\ell \cap A^2_\ell \neq \phi$. En effet, w_{i_ℓ} est identique (par définition de B_ℓ) à $w^*_{i_\ell}$, l'élément de rang ℓ d'un ordre Θ^* sans croisement pour les deux hiérarchies et $A^1_\ell \cap A^2_\ell$ contient au moins l'élément $w^*_{i_{\ell+1}}$ qui fait partie des éléments les plus proches de $w^*_{i_\ell}$ au sens de δ_1 et de δ_2 puisque les deux matrices $M(\delta_1,\Theta^*)$ et $M(\delta_2,\Theta^*)$ sont nécessairement SDDL. Si l'algorithme choisi dans $A^1_\ell \cap A^2_\ell$ un élément qui n'est pas dans $B_{\ell+1}$, il arrive nécessairement à une partie $A^1_k \cap A^2_k$ vide, avec $k > \ell$; il est donc amené à faire de nouveaux choix dans $A^1_\ell \cap A^2_\ell$ jusqu'à obtenir un élément de $B_{\ell+1}$ (un tel élément existe puisque l'intersection de $B_{\ell+1}$ et $A^1_\ell \cap A^2_\ell$ contient au moins $w^*_{i_{\ell+1}}$). On voit ainsi, en raisonnant par récurrence de $\ell = 1$ à $n-1$, que l'algorithme donne un ordre contenant les n éléments de Ω s'il existe au moins un ordre sans croisement pour les deux hiérarchies. D'après la propriété 1), cet ordre est sans croisement pour les deux hiérarchies. ∎

Exemple

On visualise deux hiérarchies indicées comme indiqué en figure 18 :

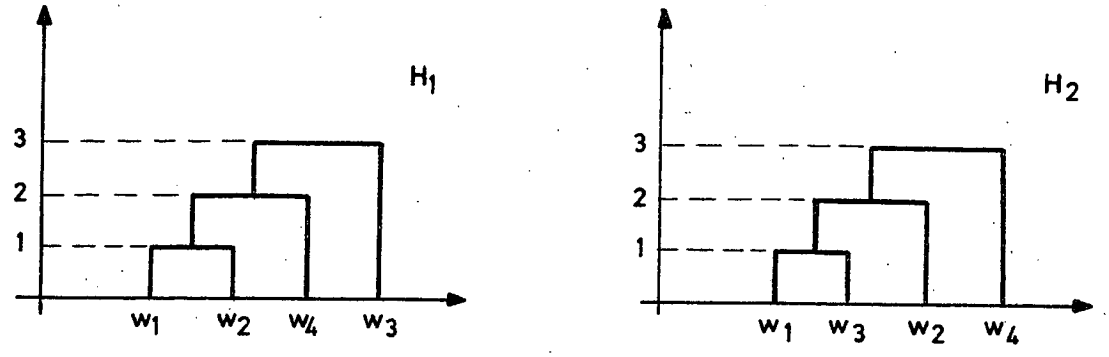


Figure 18

On remarque d'abord que $E_1 \neq E_2$ car la longueur des chaînes $w_1\ w_3\ w_2\ w_4$ et $w_1\ w_2\ w_3\ w_4$ est respectivement 6 et 8 au sens de δ_1 (voir proposition 16).

Appliquons maintenant l'algorithme sur cet exemple. On utilise, pour cela, une disposition commode qui est indiquée dans la figure 19 :

	$A^1_{i_1}$	$A^1_{i_2}$	.....	$A^1_{i_\ell}$
	$A^2_{i_1}$	$A^2_{i_1}$	.....	$A^2_{i_\ell}$
w_{i_1}	$w_{i_2} \in A^1_{i_1} \cap A^2_{i_2}$	w_{i_3}	..... w_{i_ℓ}	ϕ

Figure 19

En choisissant successivement w_1, w_2, w_2 comme points de départ, on obtient les résultats de la figure 20.

	w_2
	w_3
w_1	ϕ

	w_1	w_2
	$w_1 w_3$	w_3
w_2	w_1	ϕ

	$w_4 w_1 w_2$	w_2	w_4
	w_1	w_2	w_4
w_3	w_1	w_2	w_4

Figure 20

Le troisième tableau donne l'ordre Θ : $w_3\ w_1\ w_2\ w_4$ qui est sans croisement simultanément pour H_1 et H_2. On a représenté, figure 21, H_1 et H_2 munis de cet ordre.

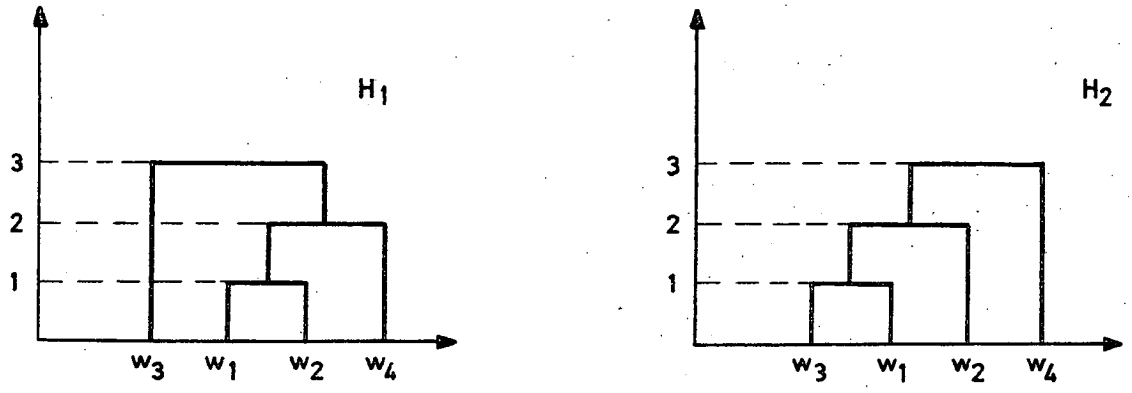


Figure 21

Remarque : Pour éliminer les éléments qui ne peuvent constituer le départ d'une chaîne de plus courte longueur (comme w_1 et w_2 dans l'exemple précédent), on peut utiliser la propriété suivante : en ordonnant les valeurs prises par la distance d'un élément w_i à tous les autres au sens de δ_ℓ, on obtient un ordre Θ_ℓ sur les couples $w_i w_\ell$: $\ell \neq i$, $1 \leq \ell \leq n$. Afin que w_i puisse être le premier élément d'une chaîne de plus courte longueur, il est nécessaire (mais non suffisant) que Θ_1 soit identique (au sens large) à Θ_2, sinon les matrices $M(\delta_1,\Theta_1)$ et $M(\delta_2,\Theta_2)$ ne pourraient être simultanément de Robinson et la chaîne associée à Θ_1 ou à Θ_2 de plus courte longueur au sens de δ_1 et δ_2.

Ainsi dans l'exemple précédent, on est assuré que w_2 est un mauvais départ car : $\delta_1(w_2,w_3) = 3 > 2 = \delta_1(w_3,w_4)$ alors que $\delta_2(w_2,w_3) = 2 < 3 = \delta(w_2,w_4)$.

Dans la pratique, il n'est pas toujours nécessaire de calculer Θ_1 et Θ_2, il suffit de détecter deux couples qui ne sont pas dans le même ordre.

14.3.- Calcul de séquences communes quand il n'existe pas d'ordre sans croisement pour les deux hiérarchies

Dans le cas où $E_1 \cap E_2 = \phi$ on peut obtenir, par le même algorithme, de grandes séquences communes en changeant l'élément de départ dès que $A^2_{i_\ell} \cap A^2_{i_\ell} = \phi$, l'ordre partiel associé à ces séquences est sans croisement d'après la proposition 17. Avec un nouvel élément, l'algorithme fonctionne comme précédemment en pensant à bien calculer les points les plus proches en tenant compte de toute la population (donc y compris les éléments qui sont dans des séquences déjà constituées), sinon la proposition 17 ne peut plus être utilisée ; cependant, pour construire la nouvelle séquence, on ne choisit quand il y a plusieurs possibilités que les éléments qui ne sont dans aucune des séquences déjà obtenues.

Exemple

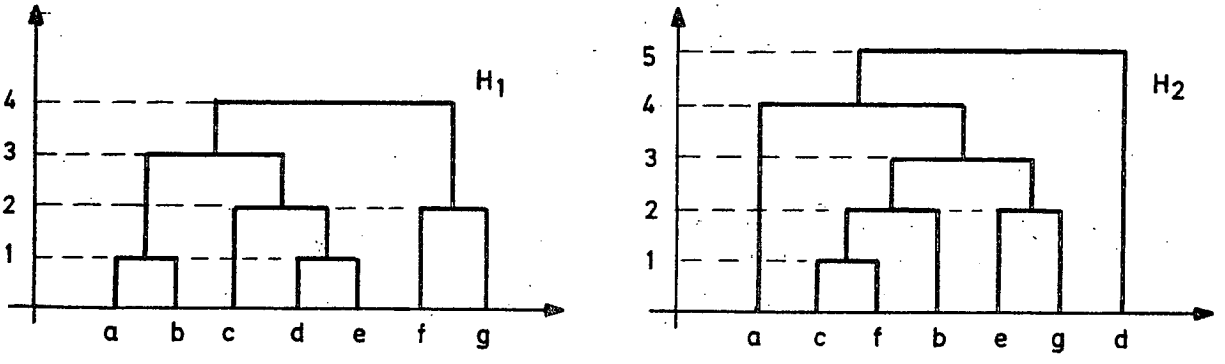


Figure 22

Les ultramétriques δ_1 et δ_2 étant induites par les hiérarchies indicées, indiquées figure 22, on obtient les 4 tableaux de la figure 23 :

	b	cde	de
	cfbeg	cf	f
a	b	c	ϕ

	e	c
	$\Omega-\{d\}$	g
d	e	ϕ

	g
	c
f	ϕ

	f
	e
g	ϕ

Figure 23

Les plus grandes séquences communes obtenues sont donc abc et de. Si l'on représente les deux hiérarchies en tenant compte de ces séquences (voir figure 24), on s'aperçoit que la méthode ne donne pas toutes les séquences communes puisque la séquence (fg) est commune.

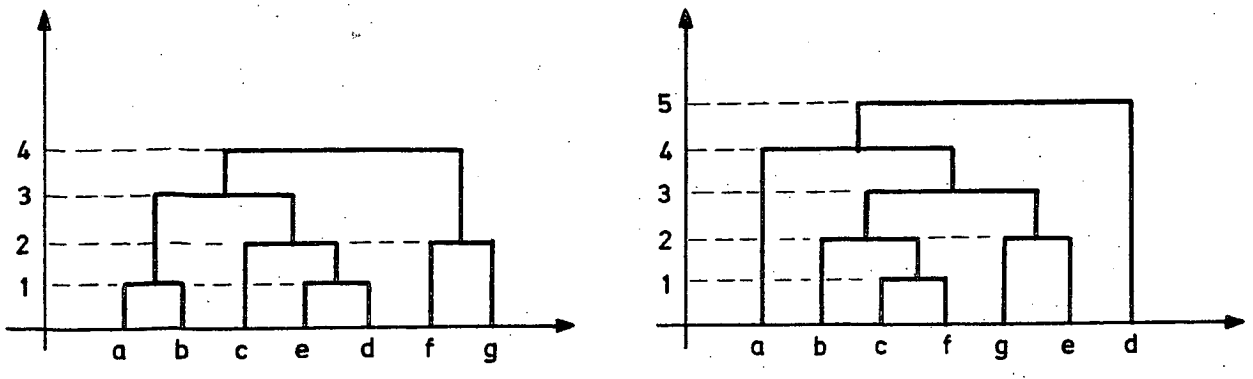


Figure 24

14.4.- Extension au cas de plusieurs hiérarchies

On peut facilement étendre tous les résultats et algorithmes au cas de plusieurs hiérarchies ; il s'agit alors de chercher un ordre sans croisement pour plusieurs hiérarchies. On étend l'algorithme du § 14.2 au cas de plusieurs hiérarchies $H_1,\ldots,H_k$, en appliquant l'algorithme du § 13 à partir d'un point commun avec les ultramétriques δ_{H_1}, ..., δ_{H_k} induites par ces hiérarchies. A chaque étape, on choisit, de façon analogue à l'algorithme du § 14.2, un nouvel élément $w_{i_{\ell+1}}$ dans $A^1_{i_\ell} \cap A^2_{i_\ell} \ldots \cap A^k_{i_k}$ où $A^j_{i_\ell}$ est l'ensemble des éléments de Ω à distance minimum de w_{i_ℓ} au sens de δ_{H_j}. Le tableau de la figure 19 se généralise sous la forme indiquée figure 25.

	$A^1_{i_1}$	.....	$A^1_{i_\ell}$
	⋮	.....	⋮
	$A^k_{i_1}$	.....	$A^k_{i_\ell}$
w_{i_1}	$w_{i_2} \in A^1_{i_1} \cap \ldots \cap A^k_{i_1}$	..... w_{i_ℓ}	ϕ

Figure 25

15.- RECHERCHE DE CONSENSUS ENTRE HIERARCHIES

Etant données plusieurs hiérarchies, il s'agit de trouver une hiérarchie qui réalise un "consensus" entre ces différentes hiérarchies (dans Rohlf (1981), on trouvera une bibliothèque récente du sujet). Il y a différentes façons d'imaginer ce "consensus" ; une façon simple étant d'utiliser la notion "d'intersection" de hiérarchies. Après avoir défini cette notion, nous donnons des algorithmes qui permettent de construire l'intersection de plusieurs hiérarchies (qui est une hiérarchie) et l'intersection du point de vue d'une hiérarchie.

15.1.- Intersections de hiérarchies

Définition

L'intersections de plusieurs hiérarchies sur Ω est l'ensemble des paliers communs à ces hiérarchies.

Proposition 18

L'intersection de plusieurs hiérarchies est une hiérarchie.

Démonstration

Considérons les hiérarchies $H_1,\ldots,H_k$ sur Ω et soit $H = H_1 \cap \ldots \cap H_k$. L'ensemble Ω et les singletons étant dans toutes ces hiérarchies sont également par définition dans l'intersection ; pour tout h_1 et h_2 dans H (c'est, par définition, un ensemble de parties de Ω), on a $h_1 \cap h_2 \neq \phi$ qui implique $h_1 \subset h_2$ ou $h_2 \subset h_1$ puisque h_1 et h_2 sont deux paliers d'une même hiérarchie H_ℓ (avec $1 \leq \ell \leq k$). H est donc une hiérarchie. ∎

Définition

L'intersection d'une hiérarchie H_i avec une hiérarchie H_j du point de vue de la hiérarchie H_i notée H_i/H_j est un ensemble de parties de Ω dont chaque élément est un palier de H_i qui est le plus petit à contenir tous les éléments d'au moins un palier de H_j.

Cette définition peut également s'énoncer de la façon suivante :

$$h \in H_i/H_j \iff \{ \exists h' \in H_j : h' \subset h \text{ avec } h \in H_i \text{ et de cardinal minimum}\}$$

A partir de cette définition, on déduit facilement la signification de $(H_1 \cap \ldots \cap H_k)/H_i$ ou (plus intéressant) de $H_i/(H_1 \cap \ldots \cap H_k)$.

Proposition 19

L'intersection de H_i avec H_j du point de vue d'une hiérarchie H_i est une hiérarchie.

Démonstration

Ω est nécessairement dans H_i/H_j puisque H_i comme H_j contient Ω . Il en est de même des singletons. Pour tout h_1 et h_2 dans H_i/H_j, $h_1 \cap h_2 = \phi$ implique $h_1 \subset h_2$ ou $h_2 \cap h_1 = \phi$ puisque h_1 et h_2 sont dans H_i qui est une hiérarchie. ∎

15.2.- Procédé constructif

On utilise l'algorithme du § 14.2 qui permet d'obtenir un ordre (quand il existe) commun à deux hiérarchies, on utilise une présentation analogue à celle de la figure 25. Les différentes notations de la figure 26 ont le sens suivant :

	$A^1_{i_1}$	.....	$A^1_{i_{n-1}}$
⋮	⋮		⋮
	$A^k_{i_1}$		$A^k_{i_{n-1}}$
w_{i_1}	w_{i_2}		w_{i_n}
$H_1 \cap \ldots \cap H_k$	h_{12}		h_{n-1n}
$H_i/(H_1 \cap \ldots \cap H_k)$	h^i_{12}		h^i_{n-1n}

Figure 26

L'ordre commun Θ : $w_{i_1}\ w_{i_2}\ \ldots\ w_{i_n}$ est obtenu comme en 14.2. $h_{\ell\ell+1}$ est par définition le plus bas des paliers communs à $H_1,\ldots,H_n$ contenant w_{i_ℓ} et $w_{i_{\ell+1}}$. On note $h^i_{\ell\ell+1}$ le plus bas palier de la hiérarchie H_i contenant le plus bas palier de la hiérarchie $H_1 \cap \ldots \cap H_k$ qui contient simultanément w_{i_ℓ} et $w_{i_{\ell+1}}$. Nous allons montrer que, pour $\ell = 1,\ldots,n-1$, l'ensemble des $h_{\ell\ell+1}$ (resp. $h^i_{\ell\ell+1}$) pour $i = 1,\ldots,k$) forme la hiérarchie $H_1 \cap \ldots \cap H_k$ (resp. $H_i/H_1 \cap \ldots \cap H_k$).

Donnons auparavant un exemple pour fixer les idées.

Exemple

Il s'agit de construire le tableau de la figure 26 à l'aide des hiérarchies H_1 et H_2 qui sont données figure 27. On obtient le tableau de la figure 28. Les hiérarchies H_1 et H_2 sont représentées selon l'ordre commun obtenu figure 29. Afin d'associer une hauteur à chaque palier des hiérarchies $H_1 \cap H_2$, H_1 / H_2 et H_2 / H_1, nous avons choisi l'indiçage suivant :

$$f_{H_1 \cap H_2} = (f_{H_1} + f_{H_2})/2$$

$$f_{H_i/H_j} = f_{H_j}$$

où $\forall h \in H$, $f_H(h)$ est la hauteur du palier h dans H.

Les hiérarchies $H_1 \cap H_2$, H_1 / H_2 et H_2 / H_1 sont données figure 30.

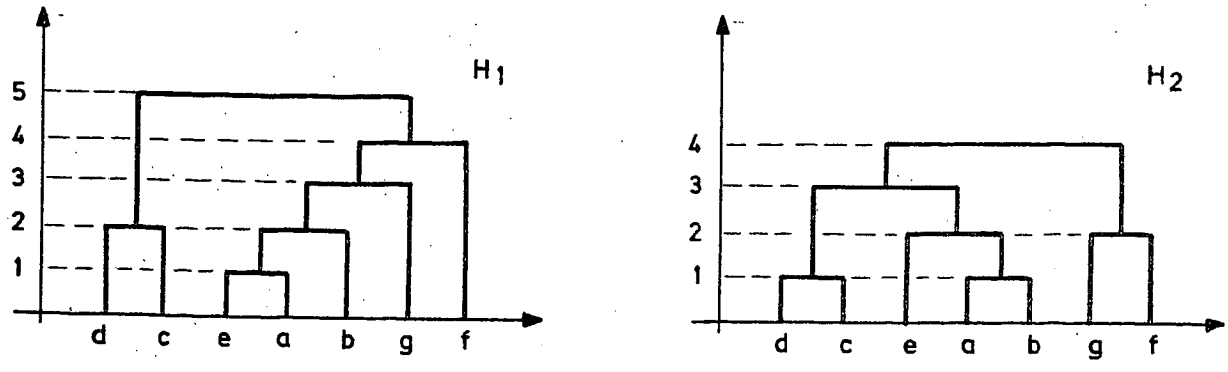


Figure 27

	eac	ea	a	b	gf	f
	c	eba	ba	b	gf	f
d	c	e	a	b	g	f
$H_1 \cap H_2$	(eacdb)	(eacdb)	(eacdb)	(eacdb)	Ω	(gf)
H_1/H_2	(eacd)	(eacdb)	(eacdb)	(eacdb)	Ω	(gf)
H_2/H_1	(eacdb)	(eacdb)	(eba)	(eacdb)	Ω	(gf)

Figure 28

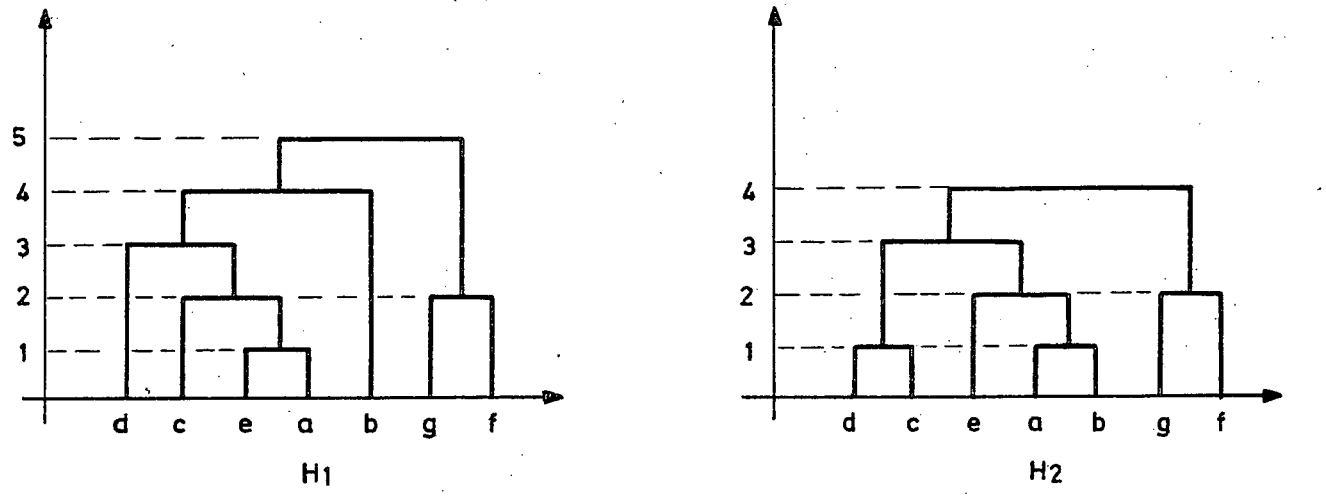


Figure 29

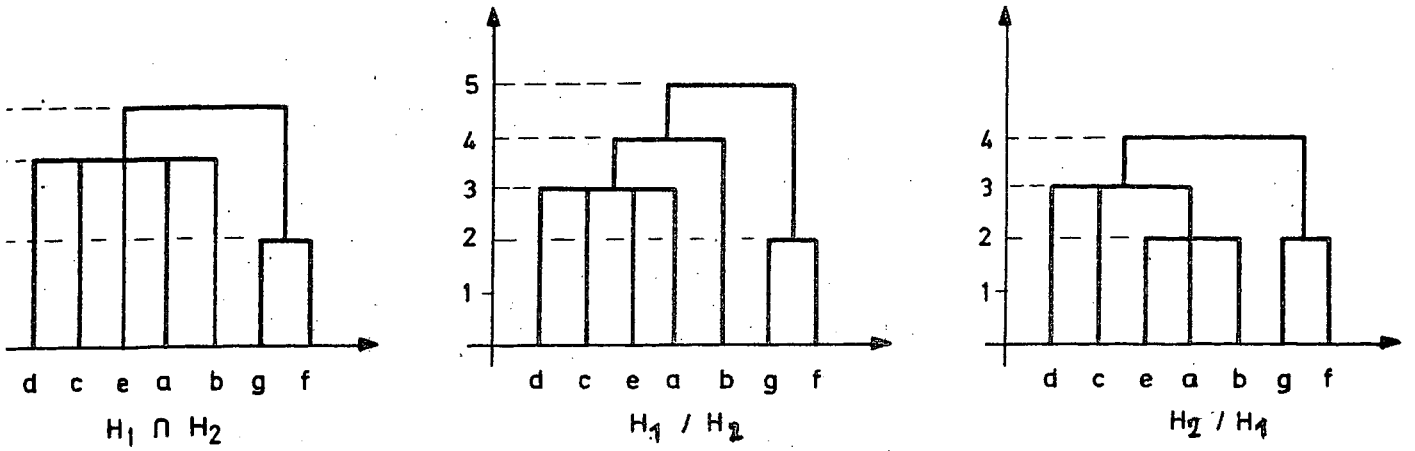


Figure 30

S'il existe un ordre commun et total Θ, on a le résultat suivant :

Proposition 20

L'ensemble des paliers $h_{\ell\ell+1}$ (resp. $h^i_{\ell\ell+1}$ pour $i = 1,\ldots,k$) pour $\ell = 1,\ldots,n-1$ définis dans le tableau de la figure 26 forme la hiérarchie $H_1 \cap \ldots \cap H_k$ (resp. $H_i/(H_1 \cap \ldots \cap H_k)$).

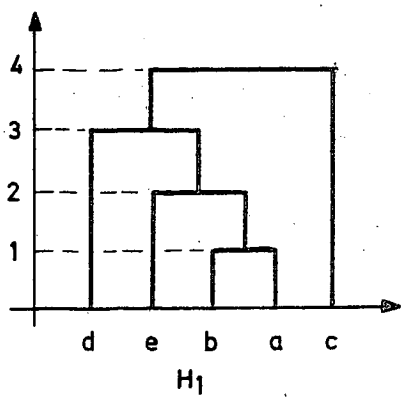
Démonstration

Pour trouver $H_1 \cap \ldots \cap H_k$, il suffit donc de comparer $h^1_{\ell\ell+1}$ pour $\ell = 1,\ldots,n-1$ à tous les paliers de H_2, ..., H_k contenant $w_{i_\ell}\ w_{i_{\ell+1}}$ et de retenir (s'il existe) celui qui est confondu à $h^1_{\ell\ell+1}$ et qui est commun à $H_2,\ldots,H_k$. De même, pour obtenir $H_i/(H_1 \cap \ldots \cap H_k)$, il suffit de comparer $h^i_{\ell\ell+1}$ pour $\ell = 1,\ldots,n-1$ au plus petit palier de $H_1 \cap \ldots \cap H_k$ qui contient $w_{i_\ell}\ w_{i_{\ell+1}}$. Si les éléments de ce palier sont dans $h^i_{\ell\ell+1}$, alors $h^i_{\ell\ell+1}$ est un palier de $H_i/(H_1 \cap \ldots \cap H_k)$. Dans les deux cas, on est assuré d'obtenir tous les paliers de $H_1 \cap \ldots \cap H_k$ et de $H_i/(H_1 \cap \ldots \cap H_k)$ puisque, en faisant varier ℓ de 1 à n-1, on atteint tous les paliers de H_1 ou de H_i (il y a une bijection entre l'ensemble des couples w_{i_ℓ} et $w_{i_{\ell+1}}$ et l'ensemble des paliers (non réduits à un singleton) si la hiérarchie est binaire , sinon c'est une surjection).

Remarque : On pourrait définir H_i/H_j comme l'ensemble (i.e. pas seulement le plus bas) des paliers de H_i qui contient les éléments d'au moins un palier de H_j. Les dernières lignes du tableau de la figure 26 seraient plus compliquées car il faudrait indiquer dans la colonne correspondante tous les paliers de H contenant les éléments du plus bas palier de H_j qui contient $w_{i_\ell}\ w_{i_{\ell+1}}$.

Exemple

Dans la figure 31, on donne deux hiérarchies H_1 et H_2. On remarque que H_1/H_2 (voir figure 32) contient le palier (eba) mais ne contient pas le palier (deba) bien que ce palier contienne tous les éléments du plus bas palier de H_2 qui contient e et a.



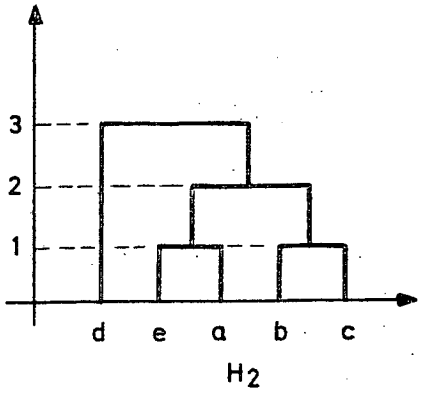


Figure 31

	eab	ba	b	c
	eabc	a	b	c
d	e	a	b	c
H_1/H_2	Ω	(eba)	Ω	Ω

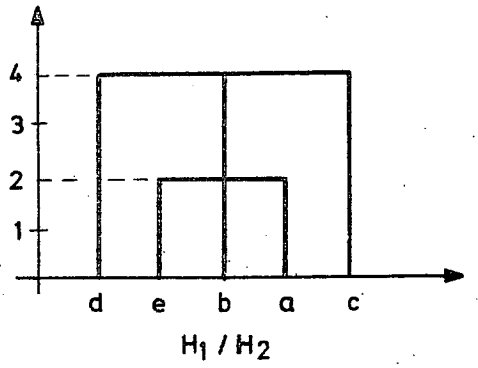


Figure 32

Si l'on utilise la définition donnée dans la remarque ci-dessus, il faut ajouter en bas de la deuxième colonne le palier (deba) ainsi que dans la hiérarchie qui est représentée figure 32.

16.- CONSTRUCTION D'UNE HIERARCHIE QUAND L'INDICE D'AGREGATION EST UNE ULTRAMETRIQUE

Le problème de la construction d'une hiérarchie à partir d'une ultramétrique se pose, par exemple, quand on cherche à minimiser un critère du type suivant :

$$\underset{\delta \in U}{\text{Min}} \quad \sum_{\substack{w_1 \in \Omega \\ w_2 \in \Omega}} [d(w_1,w_2) - \delta(w_1,w_2)]^2 \; p(w_1) \; p(w_2)$$

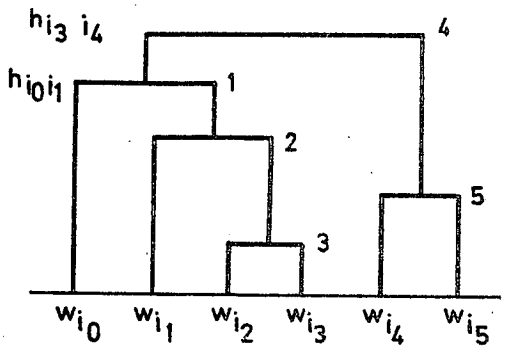
où d est la distance initiale, δ l'ultramétrique cherchée et $p(w_i)$ un poids associé à l'élément w_i.

Des algorithmes ont été proposés pour minimiser de tels critères dans Chardon et al (1978) et Fichet (1978). Ayant ainsi obtenu une ultramétrique, on utilise l'algorithme donné en 13 en associant à chaque nouvelle arête $w_{i_\ell} w_{i_{\ell+1}}$ un palier $h_{\ell\ell+1}$ à une hauteur $f(h_{\ell\ell+1}) = \delta(w_{i_\ell} w_{i_{\ell+1}})$. Ce palier contient w_{i_ℓ}, $w_{i_{\ell+1}}$ et tous les paliers avec lesquels il a une intersection non vide et qui sont à une hauteur inférieure (voir l'exemple ci-dessous). On peut montrer que la hiérarchie H ainsi construite induit une ultramétrique δ_H identique à δ.

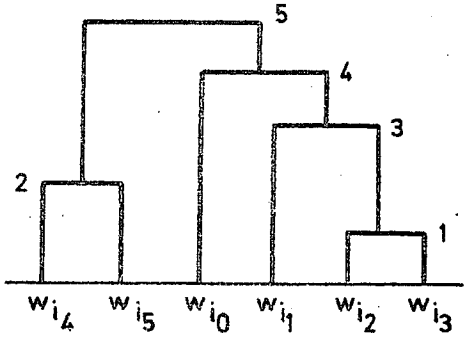
Il en résulte que cette hiérarchie ne donne pas lieu à des croisements car nous avons vu que l'algorithme donné en 13 induit une chaîne de plus courte longueur, il suffit alors d'utiliser la proposition 15.

Exemple

Dans la figure 33, on a représenté un exemple de hiérarchie construite par l'algorithme du § 13 et par l'algorithme classique de la CAH. A chaque palier est associé son numéro d'ordre dans la construction



Hiérarchie construite par l'algorithme donné au § 13.



Hiérarchie construite par l'algorithme classique de la CAH.

Figure 33

On a $h_{i_0 i_1} \cap h_{i_1 i_2} \neq \phi$ et $\delta(w_{i_0}, w_{i_1}) > \delta(w_{i_1}, w_{i_2})$ donc $h_{i_1 i_2} \subset h_{i_0 i_1}$. De même $h_{i_2 i_3} \subset h_{i_1 i_2}$; par contre, $h_{i_3 i_4} \cap h_{i_0 i_1} \neq \phi$ (puisque $h_{i_0 i_1}$ contient w_{i_3} comme $h_{i_3 i_4}$) et $\delta(w_{i_3}, w_{i_4}) > \delta(w_{i_0}, w_{i_1})$ implique que $h_{i_0 i_1} \subset h_{i_3 i_4}$.

Nous avons vu que l'algorithme donné en 13 nécessite de l'ordre de $\frac{n^2}{2}$ comparaisons. Par contre, l'algorithme classique de la CAH nécessite successivement : $\frac{n(n-1)}{2}$, $\frac{(n-1)(n-2)}{2}$, ..., 1 comparaisons, soit :

$$\frac{1}{2}(n^2 + (n-1)^2 + \ldots + 1) - \frac{1}{2}(n + (n-1) + \ldots + 1) = \frac{n(n+1)(2n+1)}{12} - \frac{n(n+1)}{4}$$

donc de l'ordre de $\frac{n^3}{6}$ comparaisons.

Remarquons enfin que l'on peut gagner aussi en utilisant l'algorithme de la CAH de façon à ne faire intervenir que les couples (puis les paliers) ordonnés suivant l'ordre obtenu par l'algorithme du § 13 puisqu'on est assuré de ne pas avoir de croisement. On réalise alors $n-1 + n-2 + \ldots + 1 = \frac{n(n-1)}{2} \sim \frac{n^2}{2}$ comparaisons ; pour construire la hiérarchie, on utilise la notation polonaise. Le nombre de comparaisons est le même que pour l'algorithme du § 13 qui a, cependant, l'avantage d'être séquentiel.

17.- CONSTRUCTION ADAPTATIVE D'UNE HIERARCHIE

17.1.- Les différents types de contraintes

Nous avons vu en 3.1 qu'il y a 2^{n-1} possibilités de représentation visuelle d'une hiérarchie ; à chacune de ces possibilités correspond un ordre différent sur les singletons. Dans la pratique, différents types de contraintes peuvent intervenir qui permettent de privilégier certains ordres.

Signalons quatre types de contraintes :

- contrainte de partition,
- contrainte d'ordre,
- contrainte de quasi-ordre,
- contrainte de hiérarchie.

La contrainte de partition apparaît quand on dispose d'une variable supplémentaire qualitative nominale. Par exemple, on a recueilli un certain nombre de mesures sur des patients et la représentation hiérarchique visuelle, faite sur le tableau de données correspondant à ces mesures, doit tenir compte au mieux du diagnostic médicale (la variable supplémentaire). Autrement dit, l'ordre des singletons doit être tel que les individus

de même diagnostic soient le moins séparés possible par des individus ayant un diagnostic différent.

La contrainte d'ordre apparaît quand on dispose d'une variable supplémentaire ordinale. L'ordre de cette variable et l'ordre des singletons de la hiérarchie à visualiser doivent être le plus "proche" possible.

La contrainte de quasi-ordre correspond au cas où la variable supplémentaire est quantitative ; en plus de l'ordre associé à cette variable, l'ordre des singletons associé à la hiérarchie doit respecter un voisinage par boule de rayon donné.

Enfin, la contrainte par hiérarchie apparaît, par exemple, quand on dispose de tableaux de données évoluant dans le temps (des images par exemple) ; il s'agit alors d'obtenir l'ordre des singletons, de la hiérarchie représentée à l'instant t+1, le plus "proche" de l'ordre des singletons de la hiérarchie correspondant à l'instant t.

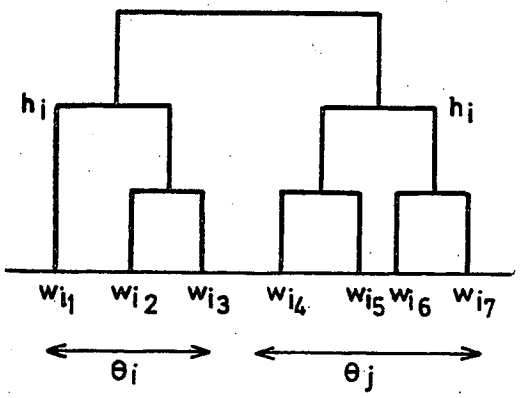
A chacune de ces contraintes, on peut associer une mesure de ressemblance d ; ainsi, dans le cas d'une contrainte de partition $d(w_i,w_j)$ égal 1 ou 0 suivant que w_i et w_j sont dans la même classe ou pas. Dans le cas d'une contrainte d'ordre $d(w_i,w_j)$ peut être le nombre d'éléments compris entre w_i et w_j. Dans le cas d'une contrainte quantitative d peut être un quasi-odre. Enfin, dans le cas d'une contrainte hiérarchie d peut être l'ultramétrique associée à cette hiérarchie.

Ainsi pour tenir compte de l'une des quatre contraintes, l'ordre des singletons de la hiérarchie à visualiser devra être modifié parmi les 2^{n-1} possibilités, de façon à tenir compte au mieux de la mesure de ressemblance associée à cette contrainte.

17.2.- Algorithmes de construction d'une hiérarchie avec contraintes

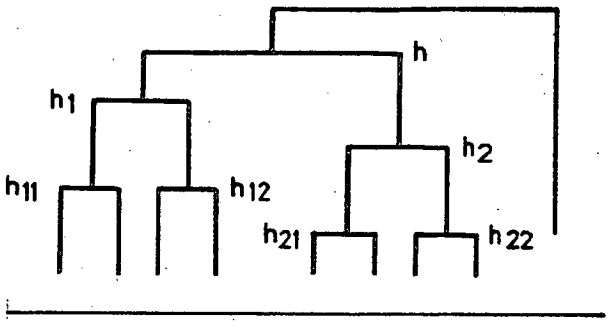
Pour construire une hiérarchie H respectant au mieux une contrainte, on se ramène donc, d'après le paragraphe précédent, à construire une chaîne de plus courte longueur, au sens de l'ultramétrique δ induite par H, "déformant" le moins possible une mesure de ressemblance d donnée.

Pour résoudre ce problème, Gravaeus et Wainer(1972) ont proposé un algorithmequi donne une solution non optimale et indépendante de l'indice d'agrégation entre classes δ_i choisi ; on considère l'ordre θ_i et θ_j associé à chaquepalier h_i et h_j tel que $h = h_i \cup h_j$; l'ordre θ, associé à h, est obtenu à partir de θ_i et θ_j en permutant les 4 extrémités de ces deux ordres de façon à obtenir l'ordre qui minimise d pour les deux extrémités les plus proches (voir figure 34). L'algorithme s'applique de bas en haut de la hiérarchie.



On permute w_{i_1} w_{i_3} w_{i_4} w_{i_7}

Algorithme de Gravaeus et Wainer



On permute h_{11} et h_{12} si $\delta_i(h_{11},h_2) < \delta_i(h_{12},h_2)$

Algorithme de Brossier

Figure 34

L'algorithme de Brossier (1980) donne un algorithme également non optimal sous la contrainte d'un indice d'agrégation δ_i entre classes choisi. Contrairement à l'algorithme de Gruvaeus et Wainer, l'algorithme de Brossier part du haut de la hiérarchie. A chaque étape, on considère le plus haut palier h non encore atteint ; soit $h = h_1 \cup h_2$ (voir figure) et $h_1 = h_{11} \cup h_{12}$, on permute h_{11} et h_{12} si $\delta_i(h_{11},h_2) < \delta_i(h_{12},h_2)$. Si $h_2 = h_{21} \cup h_{22}$, on ordonne h_{21} et h_{22} en les situant par rapport à h_{11} (si l'inégalité précédente est satisfaite).

Un troisième type d'algorithme peut être proposé en utilisant l'algorithme indiqué au § 13. On construit, à l'aide de cet algorithme, une chaîne de plus courte longueur relativement à une ultramétrique ; il doit se produire fréquemment au cours de cet algorithme, au moment de créer une nouvelle arête à partir d'une extrémité w_i de la chaîne déjà formée, que plusieurs sommets candidats soient à distance (au sens de δ) égale de cet extrémité : il suffit alors de considérer le sommet w qui minimise $d(w_i,w)$ parmi les sommets candidats (où d est la mesure de ressemblance induite par la contrainte).

Dans la pratique, l'algorithme part d'un ordre quelconque $\Theta : w_{i_1}, \ldots, w_{i_n}$ pour l'améliorer en tenant compte simultanément de deux critères :

$$W_1(\Theta) = \sum_{j=1}^{n-1} \delta(w_{i_j}, w_{i_{j+1}})$$

$$W_2(\Theta) = \sum_{j=1}^{n-1} d(w_{i_j}, w_{i_{j+1}})$$

Nous savons qu'à la convergence de l'algorithme W_1 est minimum ; pour être assuré que W_2 est également amélioré, il faut ajouter une condition supplémentaire : Partant d'un ordre quelconque $\Theta : w_{i_1}, \ldots, w_{i_n}$ l'algorithme (dans le cas général) consiste à déplacer le terme w_{i_j} pour le mettre en w_{i_ℓ} et $w_{i_{\ell+1}}$ (voir figure 35).

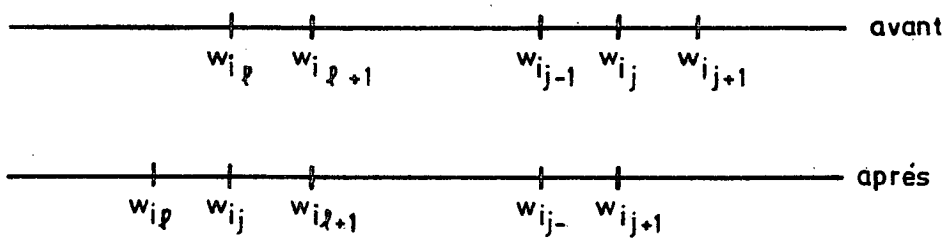


Figure 35

Pour être assuré que W_2 décroît à chaque itération, il faut vérifier que :

$$d(w_{i_\ell}, w_{i_j}) + d(w_{i_j}, w_{i_{\ell+1}}) + d(w_{i_{j-1}}, w_{i_{j+1}}) < d(w_{i_\ell}, w_{i_{\ell+1}}) + d(w_{i_{j-1}}, w_{i_j}) + d(w_{i_{j-1}}, w_{i_{j+1}})$$

Remarque : On peut utiliser de façon analogue ce troisième algorithme en remplaçant l'algorithme donné § 13 par des algorithmes proches qui permettent également d'obtenir une chaîne de plus courte longueur au sens de l'ultramétrique δ.

L'algorithme de Prim "chaîné" : si $w_i,\ldots,w_j$ est la partie de la suite déjà construite, l'étape suivante consiste à ajouter le sommet suivant w soit après w_j (comme pour l'algorithme donné en 11.1). Si $\delta(w,w_j) < \delta(w,w_i)$ soit avant w_i si $\delta(w,w_i) < \delta(w,w_j)$.

L'algorithme de Kruskal "chaîné" : on range dans l'ordre croissant les distances $\delta(w_i,w_j)$. A chaque distance $\delta(w_i,w_j)$, on associe l'arête w_iw_j. Au début la chaîne C est vide. On met dans C la première arête de la liste. On met ensuite dans la chaîne l'arête suivante sauf si un cycle peut être formé avec les crêtes qui sont déjà dans A ou si cette arête a un sommet commun avec deux arêtes de A, (c'est ici la différence avec l'algorithme classique de Kruskal). On recommence le procédé tant que la chaîne ne contient pas tous les sommets de Ω.

On montre facilement que ces deux algorithmes donnent un ordre Θ tel que la matrice $M(\delta,\Theta)$ est SDDL ; la chaîne $C(\delta,\Theta)$ est donc de plus courte longueur.

18.- CONSTRUCTION D'UNE CHAINE QUI SOIT UN ARBRE DE LONGUEUR MINIMUM

18.1.- Algorithme constructif

Si une chaîne qui soit un arbre de longueur minimum existe, elle est donnée par un algorithme de type suivant :

A l'étape 1 :

On part d'un élément quelconque $w_{i_1} \in \Omega$

A l'étape P :

On choisit w_{i_p} à partir de $w_{i_{p-1}}$ dans le complémentaire de $\{w_{i_1},\ldots,w_{i_{p-1}}\}$ parmi les éléments w' satisfaisant aux inégalités suivantes :

$$\left.\begin{array}{l} d(w',w_{i_{p-1}}) \leq d(w_{i_1},w_{i_\ell}) \\ \vdots \qquad\qquad \leq d(w_{i_2},w_{i_\ell}) \\ \vdots \qquad\qquad \vdots\ \vdots \\ \vdots \qquad\qquad \vdots\ \vdots \\ d(w',w_{i_{p-1}}) \leq d(w_{i_{p-1}},w_{i_\ell}) \end{array}\right\} \forall i_\ell \neq \{i_1,\ldots,i_{p-1}\}$$

Si w' existe, $\forall p \in \{2,\ldots,n\}$, l'ordre $\Theta : w_{i_1},\ldots,w_{i_n}$ ainsi obtenu et tel que M(d,Θ) est SDR. Si w' n'existe pas à une étape, il faut remettre en question tous les choix faits précédemment. Si malgré cette remise en question, on arrive toujours à une étape où w' n'existe pas, c'est qu'il n'existe pas de chaîne qui soit un arbre de longueur minimum.

18.2.- Déformation à donner à un indice de dissimilarité pour qu'un ordre donné induise une chaîne qui soit un arbre de longueur minimum

Si M(d,Θ) n'est pas SDR, on peut (en faisant n-1 transformations au maximum) transformer d en un indice de dissimilarité noté d_{SDR} en remplaçant chaque terme de la sur-diagonale par le plus petit terme du sous-rectangle associé.

Exemple

Dans la figure 36, on donne une matrice M(d,Θ) qui n'est pas SDR et la matrice $M(d_{SDR},\Theta)$ qui s'en déduit.

$$M(d,\Theta) = \begin{pmatrix} 0 & 3 & 1 & 2 & 3 & 6 & 5 \\ & 0 & 2 & 6 & 4 & 5 & 3 \\ & & 0 & 5 & 7 & 8 & 5 \\ & & & 0 & 6 & 4 & 5 \\ & & & & 0 & 2 & 1 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{pmatrix} \rightarrow M(d_{SDR},\Theta) = \begin{pmatrix} 0 & 1 & 1 & 2 & 3 & 6 & 5 \\ & 0 & 1 & 6 & 4 & 5 & 3 \\ & & 0 & 2 & 7 & 8 & 5 \\ & & & 0 & 3 & 4 & 5 \\ & & & & 0 & 1 & 1 \\ & & & & & 0 & 1 \\ & & & & & & 0 \end{pmatrix}$$

Figure 36

Si l'ordre Θ n'est pas donné, on peut utiliser l'algorithme donné en 18.1 et retenir par les ordres partiels ainsi obtenus celui qui contient le plus grand nombre d'éléments. On peut aussi utiliser un ordre sans croisement induit par une classification hiérarchique suivant un indice d'agrégation choisi. Rappelons, cependant, que l'indice d'agrégation du saut minimum est privilégié puisque si M(d,Θ) est SDR, l'ordre Θ est nécessairement sans croisement pour la hiérarchie du saut minimum (voir § 10).

Il serait intéressant de trouver une technique permettant de détecter parmi les 2^{n-1} ordres possibles un ordre donnant la plus petite déformation entre $M(d,\Theta)$ et $M(d_{SDR},\Theta)$.

19.- DETECTION D'ANOMALIES ENTRE DEUX HIERARCHIES INDICEES

Etant données deux hiérarchies H_1 et H_2 indicées, on peut d'abord se demander si elles sont identiques du point de vue de l'ordre des singletons Θ_1 et Θ_2 compatible avec les ultramétriques δ_1 et δ_2 induites par les hiérarchies H_1 et H_2 ; plus précisément, il s'agit de savoir si E_1 et E_2 sont identiques. Rappelons la règle simple donnée dans la proposition 16 ; il suffit que la longueur des 2 chaînes induites par Θ_1 et Θ_2 soit différente au sens de δ_1 et de δ_2 pour être assuré que $E_1 \neq E_2$.

Sachant que $E_1 \neq E_2$, on peut se demander où se situent les anomalies. Une façon simple de procéder est la suivante : on calcule la matrice des distances $M(\delta_1,\Theta_2)$, les couples correspondant à une anomalie sont ceux dont la distance associée est strictement inférieure au terme de la sur-diagonale qui se trouve sur la ligne ou la colonne correspondante.

20.- COMPARAISONS DE HIERARCHIES A L'AIDE D'UNE MESURE DE RESSEMBLANCE

De nombreux indices ont été proposés pour la comparaison de hiérarchies (voir Rohlf (1981)). On peut utiliser la notion de croisement pour évaluer "la distance" entre deux hiérarchies de la façon suivante : pour mesurer l'écart "horizontal" entre les deux hiérarchies, on peut calculer le nombre de triplets donnant lieu à un croisement entre l'ordre Θ_1 associé à la hiérarchie H_1 et la hiérarchie H_2. Notons ce nombre $\Delta(\Theta_1,H_2)$; on a alors la mesure de ressemblance suivante :

$$\Delta_{Hor}(H_1,H_2) = \frac{6(n-3)}{n!}\,(\Delta(\Theta_1,H_2) + \Delta(\Theta_2,H_1))$$

On peut aussi calculer l'écart "vertical" entre deux hiérarchies ; soit $L(H_i,\Theta_i)$ la longueur de la chaîne $C(\delta_i,\Theta_i)$ où δ_i est l'ultramétrique induite par H_i et Θ_i est un ordre sans croisement pour H_i, on en déduit la mesure de ressemblance :

$$\Delta_{Ver}(H_1,H_2) = |L(H_1,\Theta_1) - M(H_1,\Theta_2)| \cdot |L(H_2,\Theta_2) - L(H_2,\Theta_1)|$$

21.- COMPARAISON DE HIERARCHIES ET DE PARTITIONS

On peut associer une ultramétrique δ à une partition $P = (P_1,\ldots,P_k)$ de façon à ce que la hiérarchie induite par δ comporte comme seuls paliers différents de Ω et des singletons, les classes de la partition P. On pose :

$\delta(w_i,w_j) = 1$ si w_i et w_j sont dans la même classe de la partition,

$\delta(w_i,w_j) = 2$ si w_i et w_j sont dans des classes différentes de la partition.

$\delta(w_i,w_i) = 0$, $\forall i = 1, \ldots, n = \text{card } \Omega$.

Il est facile de voir que δ rend tous les triangles isocèles et d'en déduire que δ est une ultramétrique.

Il résulte de cette propriété que l'on peut utiliser tous les algorithmes précédents pour comparer des hiérarchies à des partitions.

22.- UNE DEFINITION GENERALE DE LA COMPATIBILITE DE LA SEMI-COMPATIBILITE ET DE LA COMPATIBILITE FAIBLE SUR Ω D'UN ORDRE Θ' SUR $\Omega' \subset \Omega$ ET D'UN INDICE DE DISSIMILARITE d

Afin de simplifier l'énoncé de cette définition générale le terme "X-compatibilité" est utilisé ; il peut être remplacé dans la définition par l'un des trois termes : "compatibilité", "semi-compatibilité" ou "compatibilité faible".

Définition

Un ordre Θ' : $w_{i_1} \ldots w_{i_\ell}$ (avec $\ell \leq n$) sur $\Omega' \subset \Omega$ et un indice de dissimilarité d sont dits X-compatibles sur Ω si et seulement si les deux conditions suivantes sont satisfaites :

1) $\forall w \in \Omega$, $w \notin \Omega'$ l'ordre $w\ w_{i_1} \ldots w_{i_\ell}$ ou l'ordre $w_{i_1} \ldots w_{i_\ell}\ w$ et d sont X-compatibles sur $\Omega' \cup w$; w est dit X-compatible à gauche de la chaîne $C(d,\Theta)$ dans le premier cas et X-compatible à droite dans le second.

2) Si w est X-compatible à gauche et w' est X-compatible à droite de la chaîne $C(d,\Theta)$, alors l'ordre $w\ w_{i_1} \ldots w_{i_\ell}\ w'$ est X-compatible sur $\Omega' \cup w \cup w'$.

Comme conséquences de cette définition on peut d'abord remarquer que si d et Θ sont X-compatibles alors d et Θ' sont compatibles. L'aspect matriciel conduit à la définition de trois types de matrices ; considérons une matrice $M(d,\Theta)$ où Θ est un ordre sur Ω, identique à Θ' sur Ω' ; elle fait apparaître quatre sous-matrices notées A, B, C, D. La matrice C est définie par les individus de Ω' et ses lignes et colonnes respectent l'ordre Θ'. Les colonnes de la matrice A respectent l'ordre Θ' et ses lignes sont définies par les éléments de Ω_g l'ensemble des individus à gauches de $C(d,\Theta')$. La définition des B et D se déduit immédiatement de la figure 37 où Ω_d est l'ensemble des individus situés à droite de $C(d,\Theta')$. Notons $c_1(\Theta')$ la première colonne de la matrice c et $\ell_1(\Theta')$ la dernière ligne de cette matrice.

Une matrice $M(d,\Theta,\Theta')$ telle qu'elle est définie figure 37 est pseudo-Robinson si et seulement si : les termes des lignes de la matrice A sont croissants à partir de $c_1(\Theta')$, les termes des colonnes de la matrice D sont également croissants à partir de $\ell_1(\Theta')$, enfin les termes de la matrice B sont supérieurs au terme de $c_1(\Theta')$ qui se trouve sur la même ligne et au terme de $\ell_1(\Theta')$ qui se trouve sur la même colonne.

Une matrice M(d,Θ,Θ') est pseudo-SDR si et seulement si les termes de la matrice A sont supérieurs au terme de ℓ_1 (Θ') qui se trouve sur la même colonne, les termes de la matrice D sont supérieurs au terme de c_1(Θ') qui se trouve que la même ligne, les termes de B sont supérieurs a tout terme de la sur-diagonale de C.

Une matrice M(d,Θ,Θ') est pseudo-SDD si et seulement si les termes de la matrice A (resp. D) satisfont à la même propriété que la matrice A (resp.D) d'une matrice M(d,Θ,Θ') pseudo-SDR et que les termes de la matrice B satisfont à la même propriété que la matrice B d'une matrice M(d,Θ,Θ') pseudo-Robinson. Ces trois types de matrices sont schématisées figure 38. On peut résumer l'ensemble des résultats obtenus par le schéma de la figure 39 (voir notament la proposition 8).

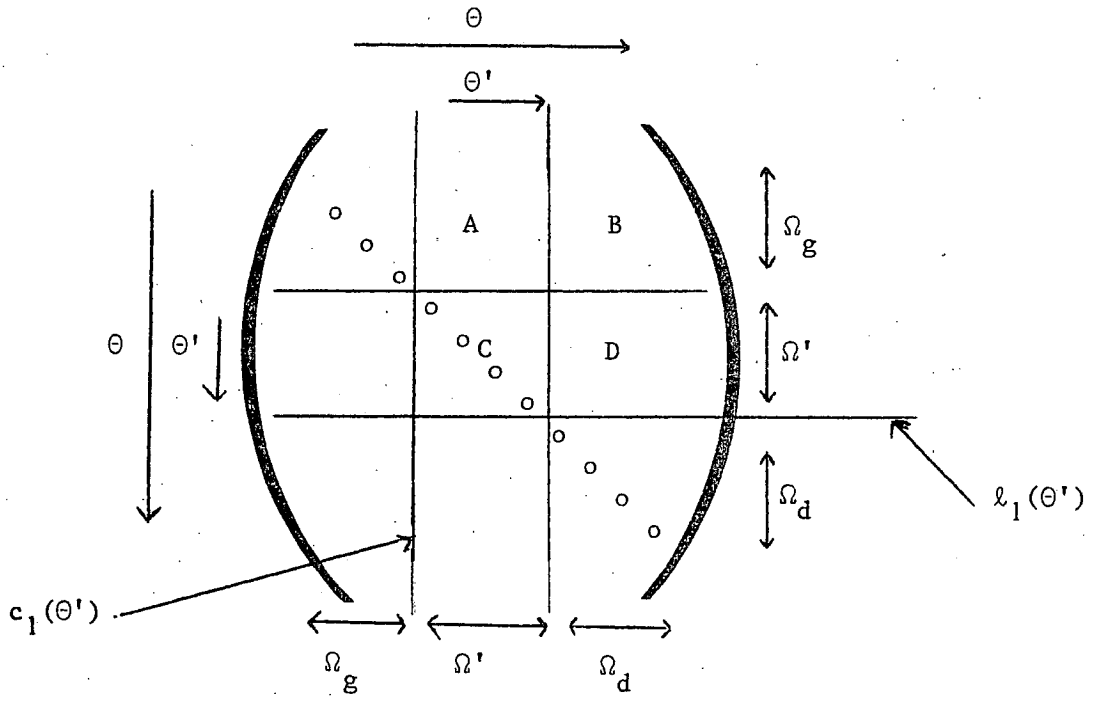
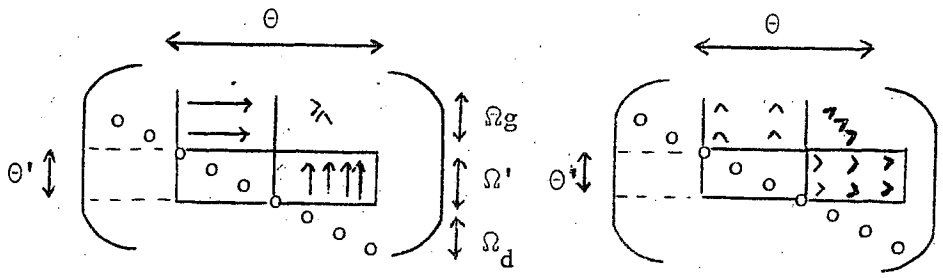


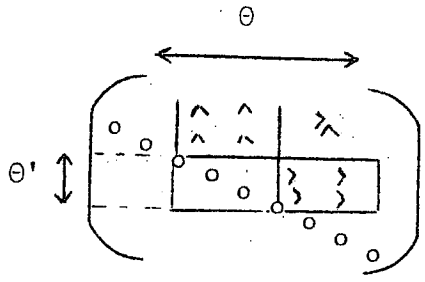
Figure 37 : Une matrice M(d,Θ,Θ')

L'ensemble des points à gauche ou à droite d'une chaîne est formé, dans le cas de la semi-compatibilité faible par exemple, de l'intersection d'hyperboloïdes plus ou moins ouvertes suivant que l'arête correspondante (coupée en son tiers) est plus ou moins longue (il faut utiliser des trapèzes à trois côtés égaux).



Pseudo-Robinson

Pseudo-SDR



Pseudo-SDD

Figure 38

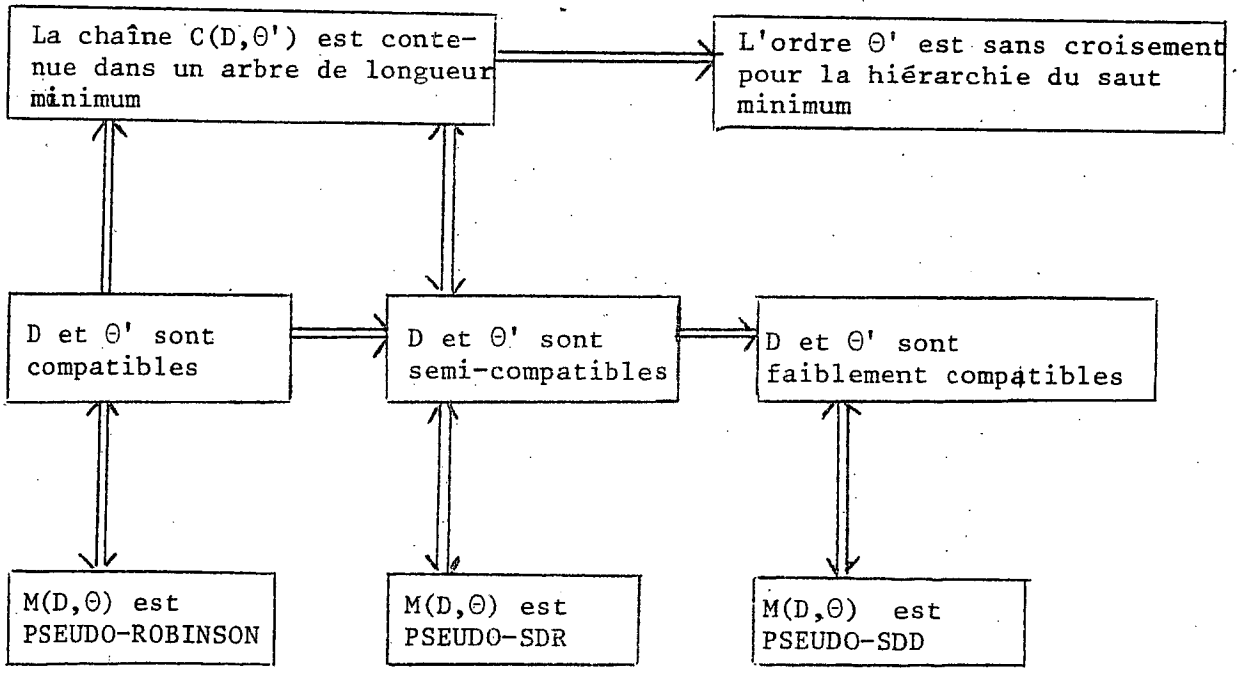


Figure 39

23 - CONCLUSION

Les résultats théoriques concernant la notion de croisement ont débouché sur une série d'algorithmes utiles pour la comparaison de hiérarchies.

La semi-compatibilité est utilisée dans le cas d'un indice de dissimilarité quelconque mais ne caractérise les ordres sans croisements que dans le cas d'une hiérarchie du saut minimimum ; par contre la compatibilité faible permet d'obtenir des ordres sans croisements dans le cas d'une hiérarchie quelconque mais nécessite la connaissance de l'ultramétrique associée.

L'introduction de nouvelles notions telles que la semi-compatibilité et la compatibilité faible ont permis de définir de nouvelles familles de matrices SDR, SDD élargissant le famille des matrices de Robinson.
On a dégagé une nouvelle caractérisation des chaînes incluses dans un arbre de longueur minimum grâce aux matrices SDR ; il en résulte un algorithme permettant de calculer de telles chaînes sans nécessiter la construction complète de l'arbre de longueur minimum. La notion de X-compatibilité d'où résultent les pseudo matrices Robinson, SDR et SDD permet d'introduire la notion de compatibilité entre un indice de dissimilarité et un ordre non total ; elle fait intervenir la notion de gauche et droite d'une chaîne permettant ainsi de situer tout objet par rapport à une chaîne.

Dans le cas d'un ordre total on a montré entre autres que la chaîne définie par un ordre sans croisement pour une hiérarchie donnée est de plus courte longueur au sens de l'ultramétrique induite par cette hiérarchie ; il en résulte un algorithme simple permettant de calculer un ordre sans croisement ; on en déduit entre autres un algorithme permettant la représentation simultanée de plusieurs hiérarchies avec le même ordre (s'il existe) sur les singletons.

De nombreuses directions de recherche restent ouvertes : Etendre les résultats obtenus au cas des populations différentes, à la comparaison d'arbres, à des cycles. Voir ce qui se passe dans le cas de distances autres que des ultramétriques (quasi-ordres, semi-ordres, etc.). Notant $\ell_c(d,\Theta)$ la longueur de la chaîne $C(d,\Theta)$, trouver l'ultramétrique δ et l'ordre Θ qui minimisent des critères du type $|\ell_c(d,\Theta) - \ell_c(\delta,\Theta)|$.

En reconnaissance des formes, utiliser la notion de "gauche" et "droite" d'une chaîne pour trouver des formes "filandreuses" ou pour situer des points par rapport à une chaîne. Il serait intéressant de voir plus précisément comment sont situés dans le plan ou dans un espace à trois dimensions les points à gauche et à droite suivant le différent type de chaîne et dans les différents cas de compatibilité entre l'ordre et la distance. Il serait aussi intéressant d'approfondir la technique de détection d'anomalies car elle peut aussi déboucher sur de nombreuses applications pratiques.

BIBLIOGRAPHIE

[1] E.N. ADAMS (1972). "Consensus techniques and the comparison of taxonomic trees", Syst. Zool, 21 : 390-397.

[2] G. BROSSIER (1980). "Représentation ordonnée des classifications hiérarchiques", Statistique et Analyse des données, vol. 2.

[3] J.L. CHANDON, J. LEMAIRE, J. POUGET. "Construction de l'ultramétrique la plus proche d'une dissimilarité au sens des moindres carrés", RAIRO, 14, 2, pp. 157-170.

[4] E. DIDAY, J. LEMAIRE, J. POUGET, F. TESTU (1982). "Eléments d'analyse des données", Dunod.

[5] E. DIDAY (1982). "Problèmes d'inversions en classification hiérarchique", Revue de Statistiques appliquées, vol. 2.

[6] J.G. FARRIS (1973)."On Comparing the shape of taxonomic trees", Syst. Zool. 22, pp. 50-54.

[7] B. FICHET (1981). "Sur des approximations d'indices de dissimilarité via les représentations euclidiennes et hiérarchiques". Revue de Statistique et Analyse des Données, Vol. 2.

[8] O. FRANK et K. SVENSSON (1981). "On probability distributions of single-linkage dendograms", J. Sat. Comput. Simul. 12, pp. 121-131.

[9] L. HUBERT (1974). "Some applications on graph theory and related non-metrics techniques to problems of approximate seriation", The British Journal of Mathematical & Statistical Psychology.

[10] L. HUBERT, F. BACKER (1977). "The comparison and Filting of given classification schemes", J. Math. Psychol. 16, pp. 233-253.

[11] T.C. HU (1960). "The maximum capacity route problem". Operation research 8, pp. 733-736.

[12] R. KALABA (1964). "Graph theory and automatic control" in : Beckenbach E.F. Ed. Applied combinatorial mathematics, New-York, Wiley.

[13] D.G. KENDALL (1969). "Incidence matrices : interval graphs and seriation in archeologic", Pacific J. Math. 28.

[14] M.F. MICKEVITCH (1978). "Taxonomic congruence", Ph. D. Dissertation State Univ. of New-York at Stony Brook, 70 pp.

[15] B. LECLERC (1974). "An application of combinatorial theory to hierarchical classification" in : BARA J.L. et Al. Eds. Recent Developments in Statistics, North Holland.

[16] B. LECLERC (1981). "Description combinatoire des ultramétriques". Math. Sc. Hum. 19ème année n° 73, pp. 5-37.

[17] I.C. LERMAN (1981). "Classification Automatique et Analyse Ordinale des Données" - DUNOD.

[18] F.J. ROHLF (1981). "Consensus Indices for Comparing Classifications". IBM Research Report R.C. 8940.

[19] P. ROSENSTHIEL (1967) -"L'arbre minimum d'un graphe" in Theorie des graphes - DUNOD.

Imprimé en France

par

l'Institut National de Recherche en Informatique et en Automatique